DOCUMENT RESUME

ED 077 285 FL 004 082

TITLE	Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, 1 July-31 December 1972.
INSTITUTION	Haskins Labs., New Haven, Conn.
SPONS AGENCY	National Inst. of Child Health and Human Development (NIH), Bethesda, Md.; National Science Foundation, Washington, D.C.; Office of Naval Research, Washington, D.C. Information Systems Research.
PUB DATE	72
NOTE	236p.
EDRS PRICE	MF-$0.65 HC-$9.87
DESCRIPTORS	Auditory Perception; Computational Linguistics; Consonants; *Electronic Equipment; *Experiments; *Information Processing; *Language Research; Memory; Neurolinguistics; Physiology; Reading Processes; *Speech; Visual Perception; Vowels

ABSTRACT

This report on speech research contains 21 papers describing research conducted on a variety of topics concerning speech perception, processing, and production. The initial two reports deal with brain function in speech; several others concern ear function, both in terms of perception and information processing. A number of reports describe electromyographic studies investigating the relationship between a particular physiological function and the production of speech sounds. Other reports deal with reading and linguistic awareness, machines and speech, and reading machines for the blind. (VM)

FILMED FROM BEST AVAILABLE COPY

ED 077285

SR-31/32 (1972)

SPEECH RESEARCH

A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for its Investigation, and Practical Applications

1 July - 31 December 1972

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the Defense Documentation Center. See the Appendix for order numbers of previous Status Reports.)

FL 004 062

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

Information Systems Branch, Office of Naval Research
Contract N00014-67-A-0129-0001

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration
Contract V101(134)P-71

National Science Foundation
Grant GS-28354

National Institutes of Health
General Research Support Grant RR-5596

National Institute of Child Health and Human Development
Contract NIH-71-2420

The Seeing Eye, Inc.
Equipment Grant

CONTENTS

I. Manuscripts and Extended Reports

The Specialization of the Language Hemisphere -- A. M. Liberman 1

A Continuum of Cerebral Dominance for Speech Perception? -- Michael Studdert-Kennedy and Donald Shankweiler 23

A Parallel Between Degree of Encodedness and the Ear Advantage: Evidence from a Temporal Order Judgment Task -- Ruth S. Day and James M. Vigorito . 41

Memory for Dichotic Pairs: Disruption of Ear Report Performance by the Speech→Nonspeech Distinction -- Ruth S. Day, James C. Bartlett, and James E. Cutting 49

Ear Advantage for Stops and Liquids in Initial and Final Position -- James E. Cutting 57

A Right-Ear Advantage in the Retention of Words Presented Monaurally -- M. T. Turvey, David Pisoni, and Joanne F. Croog 67

A Right-Ear Advantage in Choice Reaction Time to Monaurally Presented Vowels: A Pilot Study -- Michael Studdert-Kennedy 75

Perceptual Processing Time for Consonants and Vowels -- David B. Pisoni . . 83

A Preliminary Report on Six Fusions in Auditory Research -- James E. Cutting 93

Constructive Theory, Perceptual Systems, and Tacit Knowledge -- M. T. Turvey 109

Hemiretinae and Nonmonotonic Masking Functions with Overlapping Stimuli -- Claire Farley Michaels and M. T. Turvey 127

Visual Storage or Visual Masking?: An Analysis of the "Retroactive Contour Enhancement" Effect -- M. T. Turvey, Claire Farley Michaels, and Diane Kewley Port 131

Reading and the Awareness of Linguistic Segments -- Isabelle Y. Liberman, Donald Shankweiler, Bonnie Carter, and F. William Fischer 145

Machines and Speech -- Franklin S. Cooper 159

An Automated Reading Service for the Blind -- J. Gaitenby, G. Sholes, T. Rand, G. Kuhn, P. Nye, and F. Cooper 177

Physiological Aspects of Certain Laryngeal Features in Stop Production -- H. Hirose, L. Lisker, and A. S. Abramson 183

Effect of Speaking Rate on Labial Consonant Production: A Combined Electromyographic-High Speed Motion Picture Study -- Thomas Gay and Hajime Hirose 193

Stop Consonant Voicing and Pharyngeal Cavity Size -- Fredericka Bell-Berti and Hajime Hirose 207

Electromyographic Study of Speech Musculature During Lingual Nerve Block -- Gloria J. Borden, Katherine S. Harris, and Lorne Catena 213

Velar Activity in Voicing Distinctions: A Simultaneous Fiberoptic and Electromyographic Study -- Fredericka Bell-Berti and Hajime Hirose 223

Electromyographic Study of the Tones of Thai -- Donna Erickson and Arthur S. Abramson 231

II. Publications and Reports 239

III. Appendices 243

DDC and ERIC numbers (SR-21/22 - SR-29/30)

Errata (SR-28 and SR-29/30)

I. MANUSCRIPTS AND EXTENDED REPORTS

The Specialization of the Language Hemisphere*

A. M. Liberman[+]
Haskins Laboratories, New Haven

The language hemisphere may be specialized to deal with grammatical recodings, which differ in important ways from other perceptual and cognitive processes. Their special function is to make linguistic information differentially appropriate for otherwise mismatched mechanisms of storage and transmission. At the level of speech we see the special nature of a grammatical code, the special model that rationalizes it, and the special mode in which it is perceived.

The fact that language is primarily on one side of the brain implies the question I will ask in this paper: how does language differ from the processes on the other side?[1] I will suggest, as a working hypothesis, that the difference is grammatical recoding, a conversion in which information is restructured, often radically, as it moves between the sounds of speech and the messages they convey. To develop that hypothesis, I will divide it into four more specific ones: grammatical codes have a special function; they restructure information in a special way; they are unlocked by a special key; and they are associated with a special mode of perception.

---

*Invited paper presented at the Intensive Study Program in the Neurosciences (Neurosciences Research Program, Massachusetts Institute of Technology) at Boulder, Colo., July 1972.

[+]Also University of Connecticut, Storrs, and Yale University, New Haven.

[1]Language is the only cerebrally lateralized process I will be concerned with. I will not try to deal with its relation, if any, to other processes that may be in the same hemisphere, such as those underlying handedness or perception of fine temporal discriminations (Efron, 1965). Of course, we should understand cerebral specialization better if it could be shown that all the activities of one hemisphere were reflections of a single underlying design. (See, for example, Semmes, 1968.)

Acknowledgment: I am indebted to my colleagues at Haskins Laboratories, especially Franklin S. Cooper, Ignatius G. Mattingly, Donald Shankweiler, and Michael Studdert-Kennedy, for ideas, suggestions, and criticisms. Hans-Lukas Teuber, Brenda Milner, and Charles Liberman have also been very helpful. None of these people necessarily agrees with the views I express here.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

In talking about the function of grammatical codes, I will be concerned with language in general. Otherwise, I will limit my attention to speech and, even more narrowly, to speech perception. I do this partly because I know more about speech perception than about other aspects of language. But I am motivated, too, by the fact that more is known that bears on the purposes of this seminar. This becomes apparent when, in interpreting research on hemispheric specialization, we must separate processes that are truly linguistic from those that may only appear so. It becomes even more apparent when we try to frame experimental questions that might help us to discover, quite exactly, what the language hemisphere is specialized for. In any case, not so much is lost by this restriction of attention as might be supposed since, if recent arguments are accepted, speech perception is an integral and representative part of language, both functionally and formally (Liberman, 1970; Mattingly and Liberman, 1969).

THE SPECIAL FUNCTION OF GRAMMATICAL CODES: MAKING LINGUISTIC INFORMATION DIFFERENTIALLY APPROPRIATE FOR TRANSMISSION AND STORAGE

Perhaps the simplest way to appreciate the function of grammar is to consider what happens when we remember linguistic information. Should you try tomorrow to recall this lecture, we might expect, if what I say is sensible, that you would manage very well. But we can hardly conceive that you would reproduce exactly the strings of consonants and vowels, words, or sentences you will have heard. Nor can we suppose that your performance would be evaluated by any reasonable person in terms of the percentage of such elements you correctly recalled, or by the number of times your failure to recall lay merely in the substitution of a synonym for the originally uttered word. A judge of your recall would be concerned only with the extent to which you had captured the meaning of the lecture; he would expect a paraphrase, and that is what he would get.

Paraphrase is not a kind of forgetting but a normal part of remembering. It reflects the conversions that must occur if that which is communicated to us by language is to be well retained (and understood) or if that which we retain (and understand) is to be efficiently communicated. In the course of those conversions, linguistic information has at least three different shapes: an acoustic (or auditory) vehicle for transmission; a phonetic representation, consisting of consonants and vowels, appropriate for processing and storage in a short-term memory; and a semantic representation[2] that fits a nonlinguistic intellect and long-term memory. Of course, the conversions among these shapes would be of no special interest if they meant no more than the substitution of one unit for another--for example, a neural unit for an acoustic one--give or take the sharpening, distortions, and losses that must occur. But the facts of

---

[2] It may prove useful to make a distinction between a semantic representation, which presumably has linguistic structure, and some deeper base, which does not. We should suppose, then, that it is the less linguistic base that is stored in long-term memory, and that the semantic representation is synthesized from it. I believe, however, that such a distinction is not relevant to the purposes of this paper; moreover, there is no agreed-upon word to refer to the base form. I will, therefore, use "semantic representation" loosely to refer to whatever we might expect to find in long-term memory and the nonlinguistic intellect.

paraphrase imply far more than that kind of alphabetic encipherment. Since an accurate paraphrase need not, and usually does not, bear any physical resemblance to the originally presented acoustic (or auditory) signal, we must suppose that the information has been thoroughly restructured. It is as if the listener had stored a semantic representation that he synthesized or constructed out of the speech sounds, and then, on the occasion of recall, used the semantic representation as a base for synthesizing still another set of sounds. Plainly, these syntheses are not chaotic or arbitrary; they are, rather, constrained by rules of a kind that linguists call grammar. There is, therefore, a way to see the correspondence between the original and recalled information, or, indeed, between the transmitted and stored forms. But this can be done only by reference to the grammar, not by comparison of the physical properties of the two sets of acoustic events or of transforms performed directly on them. An observer who does not command the grammar cannot possibly judge the accuracy of the paraphrase.

Since my aim is to raise questions about the distinctiveness of language, I should pause here to ask whether paraphrase is unique. In visual memory, for example, is paraphrase even conceivable? Of course, the remembered scene one calls up in his mind's eye will usually differ from the original. But cannot the accuracy of recall always be judged by reference to the physical properties of the remembered scene, allowing, of course, for reversible transformations performed directly on the physical stimuli themselves? Except in the case of the most abstract art, about which there is notorious lack of agreement, can we ever say of two visual patterns that they correspond only in meaning, and, accordingly, that the correspondence between them can be judged only by reference to rules like those of grammar?

But I should return now to the function of grammatical recoding, which is the question before us. Why must the linguistic information be so thoroughly restructured if it is to be transmittable in the one case and storable in the other? The simple and possibly obvious answer is that the components for transmission and storage are grossly mismatched; consequently, they cannot deal with information in anything like the same form. I should suppose that the reason for the mismatch is that the several components developed separately in evolution and in connection with different biological activities. At the one end of the system is long-term memory, as well as the nonlinguistic aspects of meaning and thought. Surely, these must have existed before the development of language, much as they exist now in nonspeaking animals and, I dare say, in the nonlanguage hemisphere of man. At the other end of the system, the components most directly concerned with transmission--the ear and the vocal tract--had also reached a high state of development before they were incorporated as terminals in linguistic communication. [Important adaptations of the vocal tract did presumably occur in the evolution of speech, as has been shown (Lieberman, 1968, 1969; Lieberman and Crelin, 1971; Lieberman, Crelin, and Klatt, 1972; Lieberman, Klatt, and Wilson, 1969); however, these did not wholly correct the mismatch we are considering.] We might assume, then, following Mattingly (1972), that grammar developed as a special interface, joining into a single system the several components of transmission and intellect that were once quite separate. What are conceivably unique to language, to man, and to his language hemisphere are only grammatical codes. These are used to reshape semantic representations so as to make them appropriate, via a phonetic stage, for efficient transmission in acoustic form.

We should recognize, of course, that the consequences of being able to make those grammatical conversions might be immense, not merely because man can then more efficiently communicate his semantic representations to others, but also because he can, perhaps more easily than otherwise, move them around in his own head. If so, there may be thought processes that can be carried out only on information that has gone into the grammatical system, at least part way. We should also see that the nonlinguistic intellectual mechanisms might themselves have been altered in the course of evolutionary adaptations associated with the development of grammar. Indeed, exactly analogous adaptations did apparently take place at the other end of the system where, as has already been remarked, the vocal tract underwent structural changes that narrowed the gap between its repertory of shapes (and sounds) and that which was required by the nature of the phonetic representation at the next higher level. But such considerations do not alter my point, however much they may complicate it. We may reasonably suppose that the basic function of grammatical codes is to join previously independent components by making the best of what would otherwise be a bad fit.

At this point I should turn again to our question about the distinctiveness of language and ask whether the function of grammatical codes, as I described it here, is unique. Are there other biological systems in which different structures, having evolved independently, are married by a process that restructures the information passing between them? If not, then grammatical codes solve a biologically novel problem, and we should wonder whether it was in connection with such a solution that a new functional organization evolved in the left hemisphere.

But if we are to view grammar as an interface, we ought to see more clearly how bad is the fit that it corrects. For that purpose I will deal separately with two stages of the linguistic process: the interconversion between phonetic message and sound, which I will refer to throughout this paper as the "speech code," and then briefly with the part of language that lies between phonetic message and meaning.

The Phonetic Representation vs. the Ear and the Vocal Tract

At the phonetic level, language is conveyed by a small number of meaningless segments--roughly three dozen in English--called "phones" by linguists and well-known to us all as consonants and vowels. These phonetic segments are characteristic of all natural human languages and of no nonlinguistic communication systems, human or otherwise. Their role in language is an important one. When properly ordered, these few dozen segments convey the vastly greater number of semantic units; thus, they take a large step toward matching the demands of the semantic inventory to the possibilities of the vocal tract and the ear. They are important, too, because they appear to be peculiarly appropriate for storage and processing in short-term memory (Liberman, Mattingly, and Turvey, 1972). In the perception of speech the phonetic segments are retained in short-term memory and somehow organized into the larger units of words and phrases; these undergo treatment by syntactic and semantic processes, yielding up, if all goes well, something like the meaning the speaker intended. But if the larger organizations are to be achieved, the phonetic units must be collected at a reasonably high rate. (To see how important rate is, try to understand a sensible communication that is spelled to you slowly, letter by painful letter.) In fact, speaking speeds produce phonetic segments at rates of 8 to 20 segments per second, and research with artifically speeded speech (Orr, Friedman, and Williams, 1965) suggests that it is possible to perceive phonetic information at rates as high as 30 segments (that is, about seven words) per second.

Now if speech had developed from the beginning as a unitary system, we might suppose that the components would have been reasonably well matched. In that case there would have been no need for a radical restructuring of information--that is, no need for grammar--but only the fairly straightforward substitution of an acoustic segment for each phonetic one. Indeed, just that kind of substitution cipher has commonly been assumed to be an important characteristic of speech. But such a simple conversion would not work, in fact, because the requirements of phonetic communication are not directly met either by the ear or by the vocal tract.

Consider first the ear. If each phonetic unit were represented, as in an alphabet or cipher, by a unit of sound, the listener would have to identify from 8 to 30 segments per second. But such rates would surely strain, and probably overreach, the temporal resolving power of the ear. Consider next the requirement that the order of the segments be preserved. Of course, the listener would hardly be expected to order the segments if, at high rates, he could not even resolve them. We should note, however, that even at slower rates, and in cases where the identity of the sound segments is known, there is some evidence that the ear does not identify order well. Though this question has not been intensively investigated, data from the research of Warren, Obusek, Farmer, and Warren (1969) suggest that the requirements for ordering in phonetic communication would exceed the psychoacoustically determined ability of the ear by a factor of five or more.

Apparently, then, the system would not work well if the conversion from phonetic unit to sound were a simple one. We should suppose that this would be so for the reasons I just outlined. But the case need not rest on that supposition. In fact, there is a great deal of confirming evidence in the experience gained over many years through the attempts to develop and use acoustic (nonspeech) alphabets. That experience has been in telegraphy--witness Morse code, which is a cipher or alphabet as I have been using the terms here--and much more comprehensively in connection with the early attempts to build reading machines for the blind. Even after considerable practice, users do poorly with those sound alphabets, attaining rates no better than one-tenth those which are achieved in speech (Coffey, 1963; Freiburger and Murphy, 1961; Nye, 1968; Studdert-Kennedy and Cooper, 1966).

Nor does the vocal tract appear to be better suited to the requirements of phonetic communication. If the sounds of speech are to be produced by movements of the articulatory organs, we should wonder where in the vocal tract we are going to find equipment for three dozen distinctive gestures. Moreover, we should wonder, since the order of the segments must be preserved, how a succession of these gestures can be produced at rates as high as one gesture every 50 msec.

The Phonetic vs. the Semantic Representations

Though appropriate for storage over the short term, the phonetic representation apparently does not fit the requirements of the long-term store or of the essentially nonlinguistic processes that may be associated with it. Those requirements are presumably better met by the semantic representation into which the phonetic segments are converted. Because of its inaccessibility, we do not know the shape of the information at the semantic level, which is a reason we do well, for our purposes, to concentrate our attention on the acoustic

and phonetic levels where we can more readily experiment. Still, some characteristics of the semantic representation can be guessed at. Thus, given the innumerable aspects of our experience and knowledge, we should suppose that the inventory of semantic units is very large, many thousands of times larger than the two or three dozen phonetic segments that transmit it. We should suppose, further, that however the semantic units may be organized, it is hardly conceivable that they are, like the phonetic segments, set down in ordered strings. At all events, the phonetic and semantic representations must be radically different, reflecting, presumably, the differences between the requirements of the processes associated with short- and long-term memory.

THE SPECIAL RESTRUCTURING PRODUCED BY THE SPEECH CODE: SIMULTANEOUS TRANSMISSION OF INFORMATION ON THE SAME CUE

We can usefully think of grammatical coding as the restructuring of information that must occur if the mismatched components I have talked about are to work together as a single system. In developing that notion, I have so far spoken of three levels of linguistic information--semantic, phonetic, and acoustic--connected, as it were, by grammars that describe the relation between one level and the next. The phonetic and acoustic levels are linked by a grammar my colleagues and I have called the speech code. That is the grammar I shall be especially concerned with. But we should first place that grammar in the larger scheme of things, and establish some basis for demonstrating its resemblance to grammars of a more conventional kind. That has been done in some detail in recent reviews already referred to (Liberman, 1970; Mattingly and Liberman, 1969). Here I will offer the briefest possible account.

Exactly what we say about the more conventional grammars depends, of course, on which linguistic theory we choose. Fortunately, the choice is, for us, not crucial. Our purposes are well served by a very crude approximation to the transformational or generative grammar that is owing to Chomsky (1965). On that view, the conversion from semantic to phonetic levels is accomplished through two intermediate levels called "deep structure" and "surface structure." At each level--including also the phonetic, to which I have already referred--there are strings of segments (phones, words) organized into larger units (syllables, phrases). From one level to the next the organized information is restructured according to the rules of the appropriate grammar: syntax for the conversion from deep to surface, phonology for the conversion from surface to phonetic. It is not feasible to attempt an account of these grammars, even in broad terms. But I would point to one of the most general and important characteristics of the conversions they rationalize: between one level and the next there is no direct or easily determined correspondence in the number or order of the segments. Taking a simple example, we suppose that in the deep structure, the level closest to meaning, there are strings of abstract, word-like units which, when translated into the nearest kind of plain English, might say: The man is young. The man is tall. The man climbs the ladders. The ladders are shaky. According to the rules of syntax, and by taking advantage of referential identities, we should delete and rearrange the segments of the four deep sentences, emerging in due course at the surface with the single sentence: The tall young man climbs the shaky ladders. It is as if the first, second, and fourth of the deep sentences had been folded into the third, with the result that information about all four sentences is, at the surface, transmitted simultaneously and on the same words.

The information at the level of surface structure is in turn converted, often by an equally complex encoding, to the phonetic level. But I will only offer an example of one of the simplest aspects of that conversion which nevertheless shows that the information does change shape in its further descent toward the sounds of speech and also illustrates a kind of context-conditioned variation that grammatical conversions often entail. Consider in the word "ladders" the fate of the segment, spelled "s," that means "more than one." Its realization at the phonetic level depends on the segmental context: in our example, "ladders," it becomes [z]; in a word like "cats," it would be [s]; and in "house" it would be [əz].

The more obvious parts of grammar, and of the paraphrase which so strikingly reflects it, occur in the conversion between phonetic and semantic representations. But, as I have already suggested, there is another grammar, quite similar in function and in form, to be found in the speech code that connects the phonetic representation to sound. The characteristics of this code have been dealt with at some length in several recent papers (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman, Mattingly, and Cooper, 1972). I will only briefly describe some of those characteristics now to show how they might mark speech perception and, by analogy, the rest of language as different from other processes.

How the Phonetic Message is Articulated: Matching the Requirements of Phonetic Communication to the Vocal Tract

Consider, again, that there are several times more segments than there are articulatory muscle systems capable of significantly affecting the vocal output. A solution is to divide each segment into features, so that a smaller number of features produces a larger number of segments, and then to assign each feature to a significant articulatory gesture. Thus, the phonetic segment [b] is uniquely characterized by four articulatory features: stop manner, i.e., rapid movement to or from complete closure of the buccal part of the vocal tract, which [b] shares with [d, g, p, t, k, m, n, ŋ] but not with other consonants; orality, i.e., closure of the velar passage to the nose, which [b] shares with [d, g, p, t, k], but not with [m, n, ŋ]; bilabial place of production, i.e., closure at the lips, which [b] shares with [p, m] but not with [d, g, t, k, n, ŋ]; and voiced condition of voicing, i.e., vocal fold vibration beginning simultaneously with buccal opening, which [b] shares with [d, g], but not with [p, t, k].

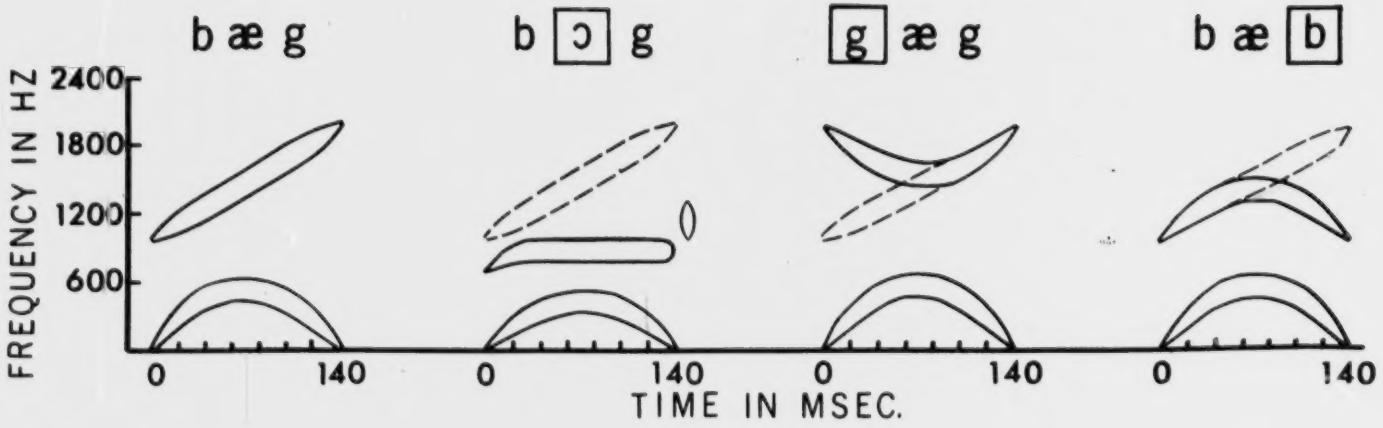
It remains, then, to produce these segments at high rates. For that purpose the segments are first organized into larger units of approximately syllabic size, with the restriction that gestures appropriate to features in successive segments be largely independent and therefore capable of being made at the same time or with a great deal of overlap. In producing the syllable, the speaker takes advantage of the possibilities for simultaneous or overlapping articulation, perhaps to the greatest extent possible. Thus, for a syllable like [bæg], for example, the speaker does not complete the lip movement appropriate for [b] before shaping the tongue for the vowel [æ] and then, only when that has been accomplished, move to a position appropriate for [g]. Rather, he overlaps the gestures, sometimes to such an extent that successive segments, or their component features, are produced simultaneously. In this way, co-articulation produces segments at rates faster than individual muscle

systems must change their states and is thus well designed, as Cooper (1966) has put it, to get fast action from relatively slow-moving machinery.

How the Co-Articulation of the Phonetic Message Produces the Peculiar Characteristics of the Speech Code

The grouping of the segments into syllables and the co-articulation of features represents an organization of the phonetic message, but not yet a very drastic encoding, since it is still possible to correlate isolable gestures with particular features. It is in the further conversions, from gestures to vocal-tract shapes to sounds, that the greater complications of the speech code are produced. For it is there that we find a very complex relation of gesture to vocal-tract shape and then, in the conversion from vocal-tract shape to sound, a reduction in the number of dimensions. The result is that the effects of several overlapped gestures are impressed on exactly the same parameter of the acoustic signal, thus producing the most important and complex characteristic of the speech code. That characteristic is illustrated in Figure 1, which is intended to demonstrate how several segments of the phonetic message are encoded into the same part of the sound. For that purpose, we begin with a simple syllable comprising the phonetic string [b] [æ] [g] and then, having shown its realization at the level of sound, we determine how the sound changes as we change the phonetic message, one segment at a time. The schematic spectrogram in the left-most position of the row at the top would, if converted to sound, produce an approximation to [bæg], which is our example. In that spectrogram the two most important formants--a formant is a concentration of acoustic energy representing a resonance of the vocal tract--are plotted as a function of time. Looking at only the second (i.e., higher) formant, so as to simplify our task, we try to locate the information about the vowel [æ]. One way to do that is to change the message from [bæg] to [bɔg] and compare the acoustic representations. The spectrogram for the new syllable [bɔg] is shown in the next position to the right, where, in order to make the comparison easier, the second formant of [bæg] is reproduced in dashed lines. Having in mind that [bæg] and [bɔg] differ only in their middle segments--that is, only in the vowels--we note that the difference between the acoustic signals is not limited, correspondingly, to their middle sections, but rather extends from the beginning of the acoustic signal to the end. We conclude, therefore, that the vowel information is everywhere in the second formant of the sound. To find the temporal extent of the [b] segment of our original syllable [bæg], we should ask, similarly, what the acoustic pattern would be if only the first segment of the phonetic message were now changed, as it would be, for example, in [gæg]. Looking, in the next position to the right, at that new syllable [gæg], we see that the change has produced a second-formant that differs from the original through approximately the first two-thirds of the temporal extent of the sound. A similar test for [g], the final consonant of our example, is developed at the right-hand end of the row; information about that segment exists in the sound over all of approximately the last two-thirds of its time course.

The general effect is illustrated in the single pattern in the lower half of the figure, which shows over what parts of the sound each of the three message segments extends. We see that there is no part of the sound that contains information about only one phonetic segment: at every point, the sound is carrying information simultaneously about at least two successive segments of the message, and there is a section in the middle where information is simultaneously



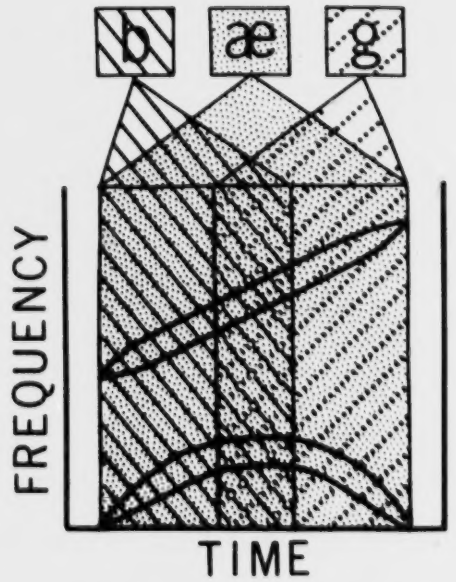


Fig. 1

Figure 1: Schemantic spectrograms showing how the segments of the phonetic message are conveyed simultaneously on the same parameter of the sound.

available about all three. It is as if the initial and final consonants [b] and [g] had been folded into the vowel, much as the flanking deep-structure sentences of the earlier syntactic example were folded into that middle sentence that served, like the vowel, as a core or carrier.

Given that information about successive segments of the message is often carried simultaneously on the same parameter of the signal, the acoustic shape of a cue for a particular segment (or feature) will necessarily be different in different contexts. To see that this is so we should look again at the figure, but instead of noting, as we did before, that changing only the middle segment caused a change in the entire acoustic signal, we should see now that, though we retained two of the three original message segments, we nevertheless left no part of the acoustic signal intact. That is to say that the acoustic cues for [b] and [g] are very different in the contexts of the different vowels into which they are encoded. Such context-conditioned variation, similar perhaps to that we noted in the phonology, is often very great, not only for a consonant segment with different vowels, as in the example offered here, but also for different positions in the syllable, different kinds of syllable boundaries, different conditions of stress, and so on. (See, for example, Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967.)

Thus, as in the conversions between other levels of the language structure, the connection between phonetic message and sound is that of a very complex code, not an alphabet or substitution cipher. Information about successive segments of the message is often encoded into a single acoustic event with the result that there is no direct or easily calculated correspondence in segmentation, and the resulting variation in the shape of the acoustic cue can be extreme. At the levels of phonetic and acoustic representations, those characteristics define what I mean by a grammatical code.

But is the speech code--and, by extension, the other grammatical codes--unique? In visual and auditory perception, the relations between stimulus pattern and perceived response may be just as complex as those of speech, but they appear, as a class, to be different. I find it difficult to characterize the difference in general terms beyond saying that, apart from speech perception, we do not find the kind of simultaneous transmission that requires the perceiver to process a unitary physical event so as to recover the two or three discrete perceptual events that are encoded in it.

How the Speech Code Matches the Requirements of Phonetic Communication to the Properties of the Ear

I remarked earlier that we can and do hear speech at rates that would appear to overreach the resolving power of the ear if each phone were transmitted by a unit sound. But we have seen that the phones are not transmitted in that direct way; they are, rather, converted so as to encode several phones into the same acoustic unit. Though this produces a great complication in the relation between signal and message, and one that will have to be dealt with by a correspondingly complex decoder, it serves the important purpose of reducing significantly the number of discrete acoustic events that must be heard, and thus makes it possible to perceive phonetic information at reasonable rates. Given that the segments are encoded into units of approximately syllabic size, we should suppose that the limit on perception is set, not by the number of phonetic segments per unit time, but more nearly by the number of syllables.

I also remarked earlier on another way in which the ear appears to be ill suited to the requirements of phonetic communication: a listener must identify the order of phonetic segments, yet in ordinary auditory perception he cannot do that well. The solution to this problem that is offered by the speech code is that order is often marked, not only by time of occurrence, but also by context-conditioned variations in the shape of the cue. Thus, because of the kind of encoding that occurs, a primary acoustic cue for the two b's in [bæb] will be mirror images of each other. In words like [tæks] and [tæsk] the acoustic cues for [k] will have very different shapes, again because of coarticulation. Hence, the speech code offers the listener the possibility of constructing (or, more exactly, reconstructing) the order of the segments out of information which is not simply, or even primarily, temporal.

More and Less Encoded Aspects of Speech

An important characteristic of the speech code, especially in relation to questions about hemispheric specialization, is that not all parts of the speech signal bear a highly encoded relation to the phonetic message. In slow to moderate articulation, vowels and fricatives, for example, are sometimes represented by a simple acoustic alphabet or cipher: there are isolable segments in which information about only one phonetic segment is carried, and there may be little variation in the shape of the acoustic cues with changes in context. Segments belonging to the classes liquids and semivowels can be said to be grammatically encoded to an intermediate degree. Though these segments cannot be isolated in the speech signal (except for r-colored vowels), they do have brief steady-state portions, even in rapid articulation.

Nongrammatical Complications in the Relation between Acoustic and Phonetic Levels

There are several characteristics of speech apart from its encodedness that might require special treatment in perception. One is that the speech signal seems very poorly designed, at least from an engineering point of view. The acoustic energy is not concentrated in the information-bearing parts of the sound but is, rather, spread quite broadly across the spectrum. Moreover, the essential acoustic cues are, from a physical point of view, among the most indeterminate. Thus, the formant transitions, so important in the perception of most consonants, are rapid changes in the frequency position of a resonance which, by their nature, scatter energy.

Another kind of difficulty arises from the gross variations in vocal tract dimensions among men, women, and children. A consequence is that the absolute values of the formant cues will be different depending on the sex and size of the speaker. Obviously, some kind of calibration is necessary if the listener is to perceive speech properly.

What, Then, Is the Language Hemisphere Specialized for?

I have suggested, as a working hypothesis, that the distinctive characteristic of language is not meaning, thought, communication, or vocalization, but, more specifically, a grammatical recoding that reshapes linguistic information so as to fit it to the several originally nonlinguistic components of the system. That hypothesis may be useful in research on hemispheric specialization for language because it tells how we might make the necessary distinction

between that which is linguistic and that which is not. Our aim, then, is to discover whether it is, in fact, the processes of grammatical recoding that the language hemisphere is specialized for. That will be hard to do at the level closest to the semantic representation because we cannot, at that end of the language system, so easily define the boundary between grammatical coding and the presumably nonlinguistic processes it serves. But in speech, and especially in speech perception, we can be quite explicit. As a result, we can ask pointed questions and, because appropriate techniques are available, get useful answers. I will offer a few examples of such questions and answers. Because the experiments I will talk about represent a large and rapidly growing class, I should emphasize that, for the special purposes of this paper, I will describe only a few.

Speech vs. nonspeech. After investigations of people with cortical lesions, including especially the studies by Milner (1954, 1958), had indicated that perception of speech and nonspeech might be primarily on opposite sides of the head, Kimura (1961a, 1961b) pioneered the development of an experimental technique that permits us to probe this possibility with normal people. Adapting for her purposes a method that had been used earlier by Broadbent (1956), Kimura presented spoken digits dichotically, one to one ear and a different one to the other ear. She discovered that most listeners heard better the digits presented to the right ear. It was subsequently found, by her and others, that the same effect is obtained with nonsense syllables, including those that differ in only one phonetic segment or feature (Kimura, 1967; Shankweiler and Studdert-Kennedy, 1967). When the stimuli are musical melodies or complex nonspeech sounds, the opposite effect, a left-ear advantage, is obtained (Kimura, 1964). On the assumption that the contralateral auditory representation is stronger than the ipsilateral, especially under conditions of dichotic competition, Kimura interpreted these findings to reflect left-hemisphere processing of the speech signals and right-hemisphere processing of the others. In any case, many studies now support the conclusion that the ear advantages are reliable reflections of the functional asymmetry of the cerebral hemispheres. (For summaries see: Kimura, 1967; Shankweiler, 1971; Studdert-Kennedy and Shankweiler, 1970.)

Auditory vs. phonetic processing. If, as seems reasonable, the right-ear advantage for speech is interpreted to reflect the work of some special device in the left hemisphere, we should ask whether that device is specialized for grammatical decoding or for something else. Consider, then, a case such as the stop consonants. As I pointed out earlier, these phonetic segments are encoded grammatically in the exact sense that there is no part of the acoustic signal that carries information only about the consonant; the formant transitions, which contain all the information about the consonant, are simultaneously providing information about the following vowel. Any device that would perceive the segments correctly must deal with that grammatical code. Conceivably, that is what the device in the language hemisphere is specialized for. But there are other, nongrammatical jobs to be done and, accordingly, other possibilities. Among these are the tasks I referred to earlier when I spoke of the need to clean up the badly smeared speech signal, to track the very rapid frequency modulations (formant transitions) that are such important cues, and to calibrate for differences in vocal-tract size. Though not grammatical according to our definition, these tasks confront the listener only in connection with speech. They might, therefore, be more closely associated with the language hemisphere than those other auditory processes that must underlie the perception of all sounds, speech and nonspeech alike. But that is precisely the kind of issue that can be settled experimentally.

Several investigators (for example Darwin, 1971; Haggard, 1971; Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy, Shankweiler, and Pisoni, 1972) have suggested and considerably refined questions like those I posed in the preceding paragraph and, in a number of ingenious experiments, found answers. I cannot here describe, or even summarize, these generally complex studies except to say that they provide some support for the notion that in speech perception the language hemisphere extracts phonetic features, which is to say in the terminology of this paper that it does grammatical decoding. There is, however, an experiment by Darwin (1971) which suggests that the language hemisphere may also be responsible for normalizing the acoustic signal to take account of the complications produced by the differences among speakers in length of the vocal tract. That finding indicates that our hypothesis is, at best, incomplete. Of course, we can hope to discover a mechanism general enough to include both vocal-tract normalization and grammatical decoding, the more so since these processes are so intimately associated with each other and with nothing else. Meanwhile, we can proceed to find out by experiment whether the language hemisphere is responsible for the other nongrammatical tasks, however closely or remotely they may be associated with speech. Perhaps an example of such an experiment will clarify the question and also our hypothesis.

Imagine a set of stop-vowel syllables [ba, da, ga] synthesized in such a way that the only distinguishing acoustic cue is the direction and extent of the first 50 msec of the second-formant transition, the rapid frequency modulation referred to earlier. Suppose, now, that we present these dichotically--that is, [ba], for example, to one ear, [da] to the other--in randomly arranged pairs and, as usual, get the right-ear advantage that is presumed to reflect left-hemisphere processing. On the hypothesis proposed here, we should say that these signals were being processed in the language hemisphere because they required grammatical decoding. In that case, we should have in the language hemisphere a device that is quite properly part of a linguistic system. There is, however, an alternative, as I have already implied, which is that the language hemisphere is specialized not for grammatical decoding but for responding to a particular class of auditory events, specifically the rapid frequency modulations of the second-formant transitions that are, in the stimuli of the experiment, the only acoustic cues. In that case, the left hemisphere would be said, at least in this respect, to be specialized for an auditory task, not a linguistic one. An experiment that helps to decide between these possibilities would go as follows. First, we remove from the synthetic syllables the second-formant transition cues and present them in isolation. When we do that we hear, not speech, but more or less distinguishable pitch glides or bird-like chirps. Now, given that there is a right-ear (left-hemisphere) advantage when these formant-transition cues are in a speech pattern, we determine the ear advantage when they are presented alone and not heard as speech. Donald Shankweiler, Ann Syrdal, and I (personal communication) have been doing that experiment. The results so far obtained are not wholly convincing, because, owing largely to the difficulty our listeners have in identifying the transition cues alone, the data are quite noisy. So far as the results can be interpreted, however, they suggest that the second-formant transitions in isolation produce a left-ear advantage, in contrast to the right-ear advantage obtained when those same transitions cued the perceived distinctions along [ba, da, ga]. If that result proves reliable, we should infer that the language hemisphere is specialized for a linguistic task of grammatical decoding, not for the auditory task of tracking formant transitions.

A clearer answer to essentially the same question, arrived at by a very different technique, is to be found in a recent doctoral dissertation by Wood (in preparation). He first replicated an earlier study (Wood, Goff, and Day, 1971) in which it had been found that evoked potentials were exactly the same in the right hemisphere whether the listener was distinguishing two syllables that differed only in a linguistically irrelevant dimension (in this case [ba] on a low pitch vs. [ba] on a high pitch) or in their phonetic identity ([ba] vs. [da] on the same pitch) but the evoked potentials in the left hemisphere were different in the two cases. From that result it had been inferred that the processing of speech required a stage beyond the processing of the nonlinquistic pitch parameter and, more important, that the stage of speech processing occurred in the left hemisphere. Now, in his dissertation, Wood has added several other conditions. Of particular interest here is one in which he measured the evoked potentials for the isolated acoustic cues, which were, as in the experiment described above, the second-formant transitions. The finding was that the isolated cue behaved just like the linguistically irrelevant pitch, not like speech. This suggests, as does the result we have so far obtained in the analogous dichotic experiment, that the processor in the language hemisphere is specialized, not for a particular class of auditory events, but for the grammatical task of decoding the auditory information so as to discover the phonetic features.

More vs. less encoded elements. As we saw earlier, only some phonetic segments--for example, [b, d, g]--are always grammatically encoded in the sense that information about them is merged at the acoustic level with information about adjacent segments. Others, such as the fricatives and the vowels, can be, and sometimes are, represented in the sound as if in a substitution cipher; that is, pieces of sounds can be isolated which carry information only about those segments. Still others, the liquids and semi-vowels, appear to have an intermediate degree of encodedness. We might suppose that only the grammatically encoded segments need to be processed by the special phonetic decoder in the left hemisphere; the others might be dealt with adequately by the auditory system. It is of special interest, then, to note the evidence from several studies that the occurrence or magnitude of the right-ear advantage does depend on encodedness (Darwin, 1971; Haggard, 1971; Shankweiler and Studdert-Kennedy, 1967). Perhaps the most telling of these experiments is a very recent one by Cutting (1972). He presented stop-liquid-vowel syllables dichotically--e.g., [kre] to one ear, [glæ] to the other--and found for the stops that almost all of his subjects had a right-ear advantage, while for the vowels the ear advantage was almost equally divided, half to the right ear and half to the left; the results with the liquids were intermediate between those extremes.

We might conclude, again tentatively, that the highly encoded aspects of speech--those aspects most in need of grammatical decoding--are always (or almost always) processed in the language hemisphere. The unencoded or less highly encoded segments may or may not be processed there. We might suppose, moreover, that some people tend to process all elements of language linguistically while others use nonlinguistic strategies wherever possible. If that is so, could it account for at least some of the individual differences in "degree" of ear advantage that turn up in almost all investigations?

Primary vs. secondary speech codes; cross codes. People ordinarily deal with the complications of the speech code without conscious awareness. But awareness of some aspects of speech, such as its phonetic structure, is sometimes

achieved. When that happens secondary codes can be created, an important example being language in its alphabetically written form. This written, secondary code is not so natural as the primary code of speech, but neither is it wholly unnatural, since it presumably makes contact with a linguistic physiology that is readily accessible when reading and writing are acquired. Research suggests that the contact is often (if not always) made at the phonetic level (Conrad, 1972); that is, that which is read is recoded into a (central) phonetic representation. If so, then we might expect to see the consequences in studies of hemispheric specialization for the perception of written language, as indeed we do (Milner, 1967; Umilta, Frost, and Hyman, 1972).

In addition to the complications of secondary codes, there are special problems arising out of the tendency, under some conditions, to cross-code nonlinguistic experience into linguistic form. As found in a recent experiment by Conrad (1972), for example, confusions in short-term memory for pictures of objects were primarily phonetic, not visual (or optical). Such results do not reveal the balance of nonlinguistic and linguistic processes, but they make it nonetheless evident that in the perception of pictures and, perhaps, of other kinds of nameable patterns, too, some aspects of the processing might be linguistic and therefore found in the language hemisphere.

A SPECIAL KEY TO THE CODE: THE GRAMMAR OF SPEECH

If the speech code were arbitrary--that is, if there were no way to make sense of the relation between signal and message--then perception could only be done by matching against stored templates. In that case there could be no very fundamental difference between speech and nonspeech, only different sets of templates. Of course, the number of templates for the perception of phonetic segments would have to be very large. It would, at the least, be larger than the number of phones because of the gross variations in acoustic shape produced by the encoding of phonetic segments in the sound; but it would also be larger than the number of syllables, because the effects of the encoding often extend across syllable boundaries and because the acoustic shape of the syllable varies with such conditions as rate of speaking and linguistic stress.

But grammatical codes are not arbitrary. There are rules--linguists call them grammars--that rationalize them. Thus, in terms of the Chomsky-like scheme I sketched earlier, the grammar of syntax tells us how we can, by rule, reshape the string of segments at the level of deep structure so as to arrive at the often very different string at the surface. In the case of the speech code we have already seen the general outlines of the grammatical key: a model of the articulatory processes by which the peculiar but entirely lawful complications of the speech code come about. The chief characteristic and greatest complication of the speech code, it will be recalled, is that information about successive segments of the message is carried simultaneously on the same acoustic parameter. To rationalize that characteristic we must understand how it is produced by the co-articulation I described earlier. Though crude and oversimple, that account of co-articulation may nevertheless have shown that a proper model of the process would explain how the phonetic message is encoded in the sound. Such a proper model would be a grammar, the grammar of speech in this case. It would differ from other grammars--for example, those of syntax and phonology--in that the grammar of speech would be a grammar done in flesh and blood, not, as in the case of syntax, a kind of algebra with no describable physiological correlates. Because the grammar of speech would correspond to an actual process,

it is tempting to suppose that the understanding of the speech code it provides is important, not just to the inquiring scientist, but also to the ordinary listener who might somehow use it to decode the complex speech sounds he hears. To yield to that temptation is to adopt what has been called a "motor theory of speech perception," and then to wonder if the language hemisphere is specialized to provide a model of the articulatory processes in terms of which the decoding calculations can be carried out.

One finds more nearly direct evidence for a motor theory when he asks which aspect of speech, articulatory movement or sound, is more closely related to its perception. That question is more sensible than might at first appear because the relation between articulation and sound can be complex in the extreme. Thus, as I have already indicated, the section of sound that carries information about a consonant is often grossly altered in different vowel contexts, though the consonant part of the articulatory gesture is not changed in any essential way. Following articulation rather than sound, the perception in all these cases is also unchanged (Liberman, 1957; Liberman, Delattre, and Cooper, 1952; Lisker, Cooper, and Liberman, 1962). Though such findings support a motor theory, I should note that only a weak form of the theory may be necessary to account for them. That is, they may only suggest that the perception of phonetic features converges, in the end, on the same neural units that normally command their articulation; in that case the rest of the processes underlying speech perception and production could be quite separate.

Evidence of a different kind can be seen in the results of a recent unpublished study by L. Taylor, B. Milner, and C. Darwin. Testing patients with excisions of the face area in the sensori-motor cortex of the left hemisphere, these investigators found severe impairments in the patients' ability to identify stop consonants (by pointing to the appropriate letter printed on a card) in nonsense-syllable contexts, though the pure-tone audiograms and performance on many other verbal tasks were normal. Patients with corresponding damage in the right hemisphere, and those with temporal or frontal damage in either hemisphere, were found not to differ from normal control subjects. It is at least interesting from the standpoint of a motor theory that lesions in the central face area did produce an inability to identify encoded stop consonants, though, as the investigators have pointed out, the exact nature of the impairment, whether of perception or of short-term memory, will be known only after further research.

The idea that the left hemisphere may be organized appropriately for motor control of articulation is in the theory of hemispheric specialization proposed by Semmes (1968). It is, perhaps, not inconsistent with her theory to suppose, as I have, that the organization of the language hemisphere makes a motor model more available to perceptual processes. But one might, on her view, more simply assume that lateralization for language arose primarily for reasons of motor control. This would fit with the suggestion by Levy (1969) that, to avoid conflict, it would be well not to have bilaterally issued commands for unilateral articulations. In that respect speech may be unique, as Evarts (personal communication) has pointed out, since other systems of coordinated movements ordinarily require different commands to corresponding muscles on the two sides. Conceivably, then, motor control of speech arose in one hemisphere in connection with special requirements like those just considered, and then everything else having to do with language followed. This assumption has the virtue of simplicity, at least in explaining how language got into one hemisphere in the first place.

Moreover, it is in keeping with a conclusion that seems to emerge from the research on patients with "split" brains, which is that, of all language functions, motor control of speech is perhaps, most thoroughly lateralized (Sperry and Gazzaniga, 1967).

At all events, though the grammar of speech makes sense of the complexly encoded relation between phonetic message and sound, it does not tell us how the decoding might be carried out. Like the other grammars of phonology and syntax, the grammar of speech works in one direction only, downward; the rules that take us from phonetic message to sound do not work in reverse. Indeed, we now know the downward-going rules well enough, at least in acoustic form, to be able to use them (via a computer) to generate intelligible speech automatically from an input of (typed) phonetic segments (Mattingly, 1968, 1971). But we do not know how to go automatically in the reverse direction, from speech sounds to phonetic message, except perhaps via the roundabout route of analysis-by-synthesis--that is, by guessing at the message, generating (by rule) the appropriate sound, and then testing for match (Stevens, 1960; Stevens and Halle, 1967).

Still, I should think that we decode speech with the aid of a model that is, in some important sense, articulatory. If so, we might suppose that the functional organization of the left hemisphere is peculiarly appropriate for the conjoining of sensory and motor processes that such a model implies.

Having said that the speech code is rationalized by a production model, I should ask whether in this respect it differs from the relations between stimulus and perception in other perceptual modalities. I think perhaps it does. In visual and auditory perception of nonverbal material the complex relations between stimulus and perception are also "ruly" rather than arbitrary, but the rules are different from those of the speech code if only because the complications between stimulus and perception in the nonspeech case do not come about as a result of the way the human perceivers produce the stimulus: the very great complications of shape constancy, for example, are rationalized, not in terms of how a perceiver makes those shapes, but by the rules of projective geometry. This is not to say that motor considerations are unimportant in nonspeech perception. Obviously, we must, in visual perception, take account of head and eye movements, else the world would appear to move when it should stand still (Teuber, 1960:1647-1648). But in those cases the motor components must be entered only as additional data to be used in arriving at the perception; the perceptual calculations themselves would be done in other terms.

I wonder, too, if the fact that the speech rules work in only one direction makes them different from those that govern other kinds of perception. In the case of shape constancy, for example, we know that one can, by the rules of geometry, calculate the image shape on the retina if he knows the shape of the stimulus object and its orientation. That would be analogous to using the grammar of speech to determine the nature of the sound, given the phonetic message. But in shape constancy, it would appear that the calculations could be made in reverse--that is, in the direction of perception. Knowing the image shape on the retina and the cues for orientation, one ought to be able to calculate directly the shape of the object. If so, then there would be no need in shape constancy, and conceivably in other kinds of nonspeech perception, for a resort to analysis-by-synthesis if the perceptual operations are to be done by calculation; rather, the calculations could be performed directly.

SPECIAL CHARACTERISTICS OF PERCEPTION IN THE SPEECH MODE

A commonplace observation about language is that it is abstract and categorical. That means, among other things, that language does not fit in any straightforward or isomorphic way onto the world it talks about. We do not use longer words for longer objects, or, less appropriately, louder words for bluer objects. If we change only one phonetic segment out of four in a word, we do not thereby create a word less different in meaning than if we had changed all four. Apart from onomatopoeia and phonetic symbolism, which are among the smallest and least typical parts of language, we do not use continuous linguistic variations to represent the continuous variations of the outside world.

It is of interest, then, to note that in the case of the encoded phonetic segments speech perception, too, is abstract. In listening to the syllable [ba], for example, one hears the stop consonant as an abstract linguistic event, quite removed from the acoustic and auditory variations that underlie it. He cannot tell that the difference between [ba] and [ga] in simplified synthetic patterns is only a rising frequency sweep in the second formant of [b] compared with a falling frequency sweep in the second formant of [g]. But if those frequency sweeps are removed from the syllable context and sounded alone, they are heard as rising and falling pitches, or as differently pitched "chirps," just as our knowledge of auditory psychophysics would lead us to expect. Perception in that auditory mode follows the stimulus in a fairly direct way; in that sense, and in contrast to the perception of speech, it is not abstract.

Perception of the encoded segments of speech is, as a corollary of its abstractness, also categorical. Thus, if we vary a sufficient acoustic cue for [b, d, g] in equal steps along a physical continuum, the listener does not hear step-wise changes but more nearly quantal jumps from one perceived category to another. This categorical perception has been measured by a variety of techniques and has been given several different but not wholly unrelated interpretations (Conway and Haggard, 1971; Fry, Abramson, Eimas, and Liberman, 1962; Fujisaki and Kawashima, 1969; Liberman, Harris, Hoffman, and Griffith, 1957; Pisoni, 1971; Stevens, Liberman, Ohman, and Studdert-Kennedy, 1969; Vinegrad, 1970). It characterizes the grammatically encoded segments (e.g., stop consonants), as I have indicated, but not the segments (e.g., the vowels in slow articulation) that are, as I noted earlier, represented in the acoustic signal as if by an alphabet or substitution cipher. Moreover, categorical perception cannot be said to be characteristic of a class of acoustic (and corresponding auditory) events, because the acoustic cues are perceived categorically only when they cue the distinctions among speech sounds; when presented in isolation and heard as nonspeech, their perception is more nearly continuous (Mattingly, Liberman, Syrdal, and Halwes, 1971).

At all events, the grammatically encoded aspects of speech do appear to be perceived in a special mode. That mode is, like the rest of language, abstract, categorical, and, perhaps more generally, nonrepresentational. Does this not present a considerable contrast to nonverbal visual and auditory perception? For all the abstracting that special detector mechanisms may do in vision or hearing, perception in those modes seems nevertheless to be more nearly isomorphic with the physical reality that occasions it. If that is truly a difference between grammatical and nongrammatical perception, it may be yet another reflection of the different organizations of the cerebral hemispheres.

SUMMARY

The aim of this paper is to suggest that the language hemisphere may be specialized to deal with grammatical coding, a conversion of information that distinguishes language from other perceptual and cognitive processes. Grammatical coding is unique, first, in terms of its function, which is to restructure information so as to make it appropriate for long-term storage and (nonlinguistic) cognitive processing at the one end of the system and for transmission via the vocal tract and the ear at the other.

To see further how grammatical restructurings are unique, we should look, more narrowly, at the speech code, the connection between phonetic message and sound. There we see a grammatical conversion that produces a special relation between acoustic stimulus and perception: information about successive segments of the perceived phonetic message is transmitted simultaneously on the same parameter of the sound. On that basis we can tentatively distinguish that which is grammatical or linguistic from that which is not. Then, by taking advantage of recently developed experimental techniques, we can discover to what extent our hypothesis about hemispheric specialization is correct and how it needs to be modified.

The speech code is unique in still other ways that may be correlates of the special processes of the language hemisphere. Thus, the speech code requires a special key. To understand the relation between acoustic stimulus and perceived phonetic message, one must take account of the manner in which the sound was produced. Conceivably, the language hemisphere is specialized to provide that "understanding" by making available to the listener the appropriate articulatory model.

The speech code is unique, too, in that it is associated with a special mode of perception. In that mode perception is categorical, digital, and most generally, nonrepresentational. Perhaps these perceptual properties reflect the specialized processes of the language hemisphere.

REFERENCES

Broadbent, D. E. (1956) Successive responses to simultaneous stimuli. Quart. J. Exp. Psychol. 8, 145-162.

Chomsky, N. (1965) Aspects of the Theory of Syntax. (Cambridge, Mass.: MIT Press).

Coffey, J. L. (1963) The development and evaluation of the Batelle aural reading device. Proceedings of the International Congress on Technology and Blindness I. (New York: American Foundation for the Blind) 343-360.

Conrad, R. (1972) Speech and reading. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 205-240.

Conway, D. A. and M. P. Haggard. (1971) New demonstrations of categorical perception. In Speech Synthesis and Perception Progress Report No. 5. (Cambridge, England: Psychological Laboratory) 51-73.

Cooper, F. S. (1966) Describing the speech process in motor command terms. J. Acoust. Soc. Amer. 39, 1121-A. (Also in Haskins Laboratories Status Report on Speech Research SR-5/6, 2.1-2.27.)

Cutting, J. E. (1972) A parallel between encodedness and the magnitude of the right ear effect. Haskins Laboratories Status Report on Speech Research SR-29/30, 61-68.

Darwin, C. J. (1971) Ear differences in recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 46-62.
Efron, R. (1965) The effect of handedness on the perception of simultaneity and temporal order. Brain 86, 261-284.
Freiberger, J. and E. G. Murphy. (1961) Reading machines for the blind. IRE Professional Group on Human Factors in Electronics HFE-2, 8-19.
Fry, D. B., A. S. Abramson, P. D. Eimas, and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Lang. Speech 5, 171-189.
Fujisaki, H. and T. Kawashima. (1969) On the modes and mechanisms of speech perception. In Annual Report No. 1. (Tokyo: University of Tokyo, Division of Electrical Engineering, Engineering Research Institute) 67-73.
Haggard, M. P. (1971) Encoding and the REA for speech signals. Quart. J. Exp. Psychol. 23, 34-45.
Kimura, D. (1961a) Some effects of temporal-lobe damage on auditory perception. Canad. J. Psychol. 15, 156-165.
Kimura, D. (1961b) Cerebral dominance and perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.
Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exp. Psychol. 16, 355-358.
Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.
Levy, J. (1969) Possible basis for the evolution of lateral specialization of the human brain. Nature 224, 614-615.
Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Amer. 29, 117-123.
Liberman, A. M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.
Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1952) The role of selected stimulus variables in the perception of the unvoiced stop consonants. Amer. J. Psychol. 65, 497-516.
Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.
Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington, D. C.: V. H. Winston) 307-334.
Lieberman, P. (1968) Primate vocalizations and human linguistic ability. J. Acoust. Soc. Amer. 44, 1574-1584.
Lieberman, P. (1969) On the acoustic analysis of primate vocalizations. Behav. Res. Meth. Instrum. 1, 169-174.
Lieberman, P. and E. S. Crelin. (1971) On the speech of Neanderthal man. Ling. Inq. 2, 203-222.
Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man and chimpanzee. American Anthropologist 74, 287-307.
Lieberman, P., D. H. Klatt, and W. A. Wilson. (1969) Vocal tract limitations of the vocal repertoires of Rhesus monkey and other nonhuman primates. Science 164, 1185-1187.
Lisker, L., F. S. Cooper, and A. M. Liberman. (1962) The uses of experiment in language description. Word 18, 83-106.
Mattingly, I. G. (1968) Synthesis by rule of General American English. Supplement to Haskins Laboratories Status Report on Speech Research, 1-223.

Mattingly, I. G. (1971) Synthesis by rule as a tool for phonological research. Lang. Speech 14, 47-56.

Mattingly, I. G. (1972) Speech cues and sign stimuli. American Scientist 60, 327-337.

Mattingly, I. G. and A. M. Liberman. (1969) The speech code and the physiology of language. In Information Processing in the Nervous System, ed. by K. N. Leibovic. (New York: Springer Verlag) 97-117.

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Milner, B. (1954) Intellectual functions of the temporal lobe. Psychol. Bull. 51, 42-62.

Milner, B. (1958) Psychological defects produced by temporal lobe excision. Proceedings of the Association for Research of Nervous and Mental Disorders 36, 244-257.

Milner, B. (1967) Brain mechanisms suggested by studies of temporal lobes. In Brain Mechanisms Underlying Speech and Language, ed. by F. L. Darley. (New York: Grune and Stratton) 122-132.

Nye, P. (1968) Research on reading aids for the blind--a dilemma. Med. Biolog. Eng. 6, 43-51.

Orr, D. B., H. L. Friedman, and J. C. C. Williams. (1965) Trainability of listening comprehension of speeded discourse. J. Educ. Psychol. 56, 148-156.

Pisoni, D. (1971) On the nature of categorical perception of speech sounds. Doctoral dissertation, University of Michigan. (Reproduced as Supplement to Haskins Laboratories Status Report on Speech Research.)

Semmes, J. (1968) Hemispheric specialization: A possible clue to mechanism. Neuropsychologia 6, 11-27.

Shankweiler, D. (1971) An analysis of laterality effects in speech perception. In The Perception of Language, ed. by D. L. Horton and J. J. Jenkins. (Columbus, Ohio: Chas. E. Merrill) 185-200.

Shankweiler, D. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Sperry, R. W. and M. S. Gazzaniga. (1967) Language following surgical disconnection of the hemispheres. In Brain Mechanisms Underlying Speech and Language, ed. by C. H. Millikan and F. L. Darley. (New York: Grune and Stratton) 108-121.

Stevens, K. N. (1960) Toward a model for speech recognition. J. Acoust. Soc. Amer. 32, 47-55.

Stevens, K. N. and M. Halle. (1967) Remarks on analysis by synthesis and distinctive features. In Models for the Perception of Speech and Visual Form, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press) 88-102.

Stevens, K. N., A. M. Liberman, S. E. G. Ohman, and M. Studdert-Kennedy. (1969) Cross-language study of vowel perception. Lang. Speech 12, 1-23.

Studdert-Kennedy, M. and F. S. Cooper. (1966) High-performance reading machines for the blind; psychological problems, technological problems, and status. Proceedings of Saint Dunstan's Conference on Sensory Devices for the Blind. (London) 317-342.

Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Studdert-Kennedy, M., D. Shankweiler, and D. Pisoni. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. Cog. Psychol. 3, 455-466.

Teuber, H.-L. (1960) Perception. In Handbook of Physiology Section 1: Neurophysiology, Volume III, ed. by J. Field. (Washington, D. C.: American Physiological Society) 1595-1669.

Umilta, C., N. Frost, and R. Hyman. (1972) Interhemispheric effects on choice reaction times to one-, two-, and three-letter displays. J. Exp. Psychol. 93, 198-204.

Vinegrad, M. (1970) A direct magnitude scaling method of investigating categorical versus continuous modes of speech perception. Haskins Laboratories Status Report on Speech Research SR-21/22, 147-156.

Warren, R. M., C. J. Obusek, R. M. Farmer, and R. T. Warren. (1969) Auditory sequence: Confusion of patterns other than speech or music. Science 164, 586-587.

Wood, C. C. (in preparation) Levels of processing in speech perception: Neurophysiological and cognitive analyses. Unpublished Ph.D. dissertation, Yale University.

Wood, C. C., W. R. Goff, and R. S. Day. (1971) Auditory evoked potentials during speech perception. Science 173, 1248-1251.

A Continuum of Cerebral Dominance for Speech Perception?*

Michael Studdert-Kennedy[+] and Donald Shankweiler[++]
Haskins Laboratories, New Haven

ABSTRACT

A group of 22 unselected adults and a group of 30 right-handed male adults were tested on a series of handedness measures and on a dichotic CV-syllable test. Multiple regression methods were used to determine a correlation coefficient between handedness measures and dichotic ear advantages of .69 ($p < .05$) for the first group and of .54 ($p < .01$) for the second group. Implications of these findings for the concept of cerebral dominance are discussed.

Cerebral dominance for language is commonly treated as a discrete two-, or at most three-, valued variable. This is largely due to the nature of the observations that support the concept and its operational definition. For as Semmes (1968:11) has remarked, "...the concept is little more than a label, a restatement of the findings that lesions of one hemisphere produce deficits that lesions of the other hemisphere do not." Nonetheless, the suspicion that individuals may vary in their degree of hemispheric asymmetry has been repeatedly expressed in the literature (e.g., Zangwill, 1960; Hecaen and Ajuriaguerra, 1964). Often the suspicion arises in discussion of left-handed individuals in whom the severity and duration of aphasia tends to be reduced. For such cases "greater hemispheric equipotentiality" may be hypothesized (Subirana, 1958) and the intra-carotid sodium amytal test has provided direct evidence of this: some left-handers display disturbance of speech upon injection of either hemisphere (Milner, Branch, and Rasmussen, 1966). Luria (1966) has extended the hypothesis to include right-handed individuals. From observations of some 800 patients he concludes that individual differences in degree of aphasic disturbance "cannot be entirely explained by the severity of the lesion....The degree of dominance of one hemisphere in relation to lateralized processes such as speech varies considerably from case to case" (p. 89).

---

*Revised version of a paper read before the Academy of Aphasia, Rochester, N. Y., October 1972, by M. Studdert-Kennedy.

[+]Also Queens College and the Graduate Center of the City College of New York.

[++]Also University of Connecticut, Storrs.

Acknowledgment: We thank Nina deJongh for devising the scissors and tracing tasks, and for collecting and analyzing the data; and Gary Kuhn for writing the computer programs used in data analysis.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

While there may be no reason to doubt the generality of Luria's conclusions, they were necessarily reached by relatively coarse, ordinal measurement of aphasic disturbance in an arduously accumulated population of patients. The advent and refinement of the dichotic technique developed by Kimura (1961a, 1961b) have made it possible to test Luria's hypothesis on normal subjects. As known, subjects asked to recognize dichotically presented speech sounds tend to perform better on sounds presented to their right ears. Kimura (1961a, 1961b, 1967) has hypothesized that this right-ear advantage reflects the greater efficiency of the contralateral pathway, under conditions of dichotic competition, and dominance of the left hemisphere for language functions. Her own and others' work have by now amply supported this interpretation.

However, one aspect of the ear advantages deserves more experimental attention: individuals differ quite widely in the size and direction of their ear advantages. Variations in direction (left ear/right ear) are almost certainly associated with variations in the language dominant hemisphere. Kimura (1961b) found that patients, known by sodium amytal test to have speech represented in the left hemisphere, were more accurate in reporting dichotic speech sounds presented to their right ears, while patients known to have right hemisphere speech representation were more accurate on those presented to their left ears. Furthermore, groups of left-handed subjects [among whom there is likely to be a fair number of individuals having speech represented in the right hemisphere (Milner, Branch, and Rasmussen, 1966)] show reduced mean right-ear advantages or mean left-ear advantages (Bryden, 1965, 1970; Curry, 1967; Satz, Achenbach, Pattishall, and Fennel, 1965; Zurif and Bryden, 1969).

But variations in the <u>size</u> of the ear advantage within homogeneous handedness groups are more puzzling. As was earlier remarked, two conditions are presumed necessary for an ear advantage to occur in dichotic studies: greater efficiency of the contralateral pathway and cerebral dominance for language. To which of these sources is the variability in ear advantages to be attributed? To both? To neither?

Here two facts may serve us in good stead. First is the known relation between handedness and cerebral dominance for language. Second is the fact that handedness may be measured reliably along a continuum (Benton, Myers, and Polder, 1962; Benton, 1965; Satz, Achenbach, and Fennel, 1967; Annett, 1970). For if some portion of the variability in ear advantage is, indeed, due to variations in the degree of cerebral dominance, we should expect to find a significant correlation between ear advantages and continuous measures of handedness. A significant association between these variables has, in fact, been reported by Satz and his colleagues (Satz, Achenbach, and Fennel, 1967). They showed that the association increased if handedness measures were used to reclassify self-classified left- and, to some extent, right-handers. Our approach, in contrast, is to scrap the categories, to treat both handedness and dichotic ear advantage as continuous variables, and to measure the correlation between them.

For children this correlation has already been demonstrated. Orlando (1971) used a dichotic consonants test (of the type used in the present study) on 4th and 6th grade boys. He found a significant correlation between ear advantages and scores on a battery of dexterity tests, for both right- and left-handed groups. However, the subjects of these experiments were children for

whom both dominance and handedness may still have been in the process of development. The present report is a preliminary account of a study extending the method to adults for whom dominance and handedness may be presumed stable.

METHOD

Subjects

Results are reported here for two groups of subjects screened for normal hearing by audiometry. Group 1 consists of 22 unselected adults, including 4 right-handed and 1 left-handed female, 14 right-handed and 3 left-handed males. Group 2 consists of the 14 right-handed males of Group 1 together with 16 other right-handed males, added later to make a total of 30. Handedness classification is here based on answers to the six "primary questions" of Annett (1970): with which hand do you write, throw a ball, swing a racket, strike a match, hammer a nail, brush your teeth? A subject was classified as right- or left-handed only if he answered all six questions consistently.

Subjects were run in a dozen or so one-hour sessions distributed over roughly two weeks. They were tested individually in a series of handedness tasks on the first and last days. On the intervening days they were tested in groups of 4 on a series of dichotic listening tasks. They were paid for their work.

Handedness Tasks

Subjects were asked to perform on seven handedness tests, assessing three aspects of handedness (speed, strength, dexterity) that may or may not be related. They performed each task once on the first day and once on the last day. The order of the first six tasks was different for each subject and was reversed on the second run. The seventh task (strength of grip) was taken last on each day by all subjects. Hand order was counterbalanced within a subject beginning with the preferred hand. A list and brief description of the tasks for each hand on a single day follows:

1. Scissors. Time in seconds to cut a complex shape accurately (see Figure 1).
2. Tracing. Time in seconds to trace accurately a complex pattern between parallel lines 1 mm apart (see Figure 2).
3. Crawford Screws [a subtest of the Small Parts Dexterity Test (Crawford and Crawford, 1956)]. Number of small screws inserted by one hand, with support from the other, in 2 min.
4. Crawford Pegs (a subtest of the Small Parts Dexterity Test). Number of pegs inserted and washers mounted by one hand, with tweezers, in 2 min.
5. Tapping. Number of taps with metal stylus on metal plate, counted electrically over six 15-sec trials.
6. Purdue Pegboard. Number of pegs placed in a row over two 30-sec trials.
7. Stoelting Dynamometer. Total kilograms of pull on three trials.

The test-retest reliabilities of the last two tasks were less than .30 for the first group. Accordingly, only the first five (Scissors, Tracing, Crawford Pegs, Crawford Screws, and Tapping) were used for later analysis.

SCISSORS PATTERN
(Measure: time in seconds to complete)

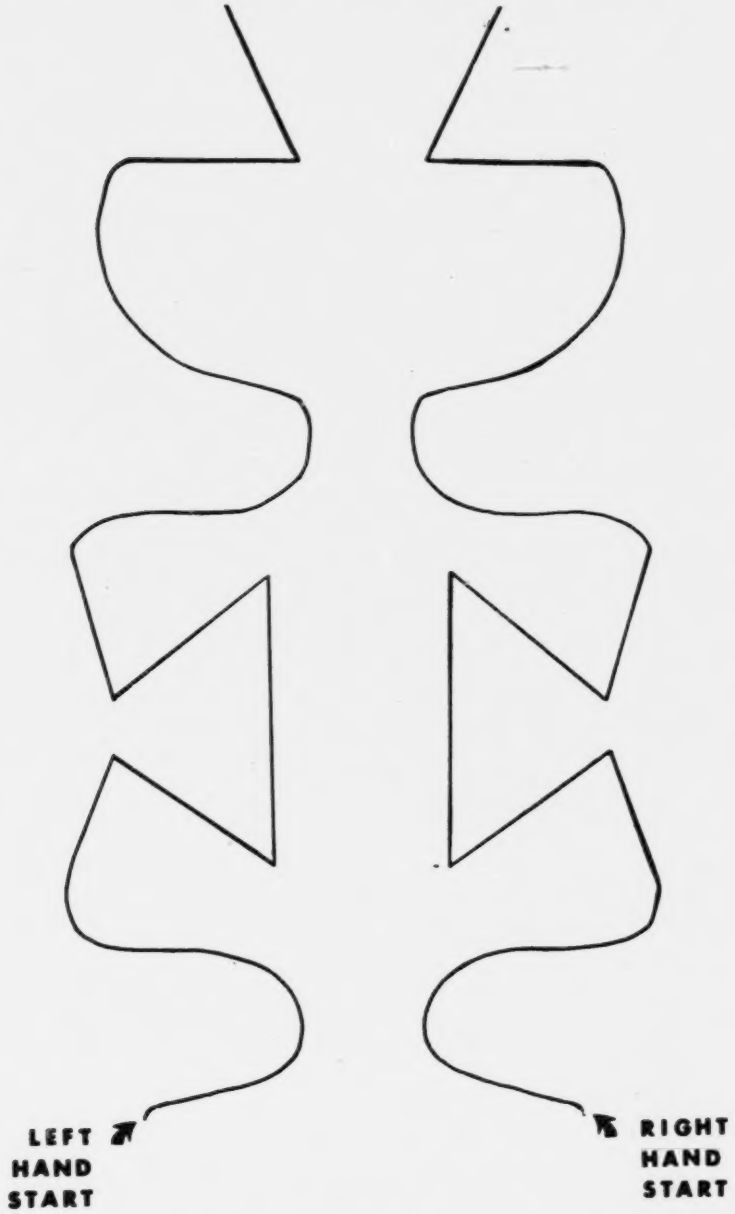


Fig. 1

TRACING PATTERN
(Measure: time in seconds to complete)

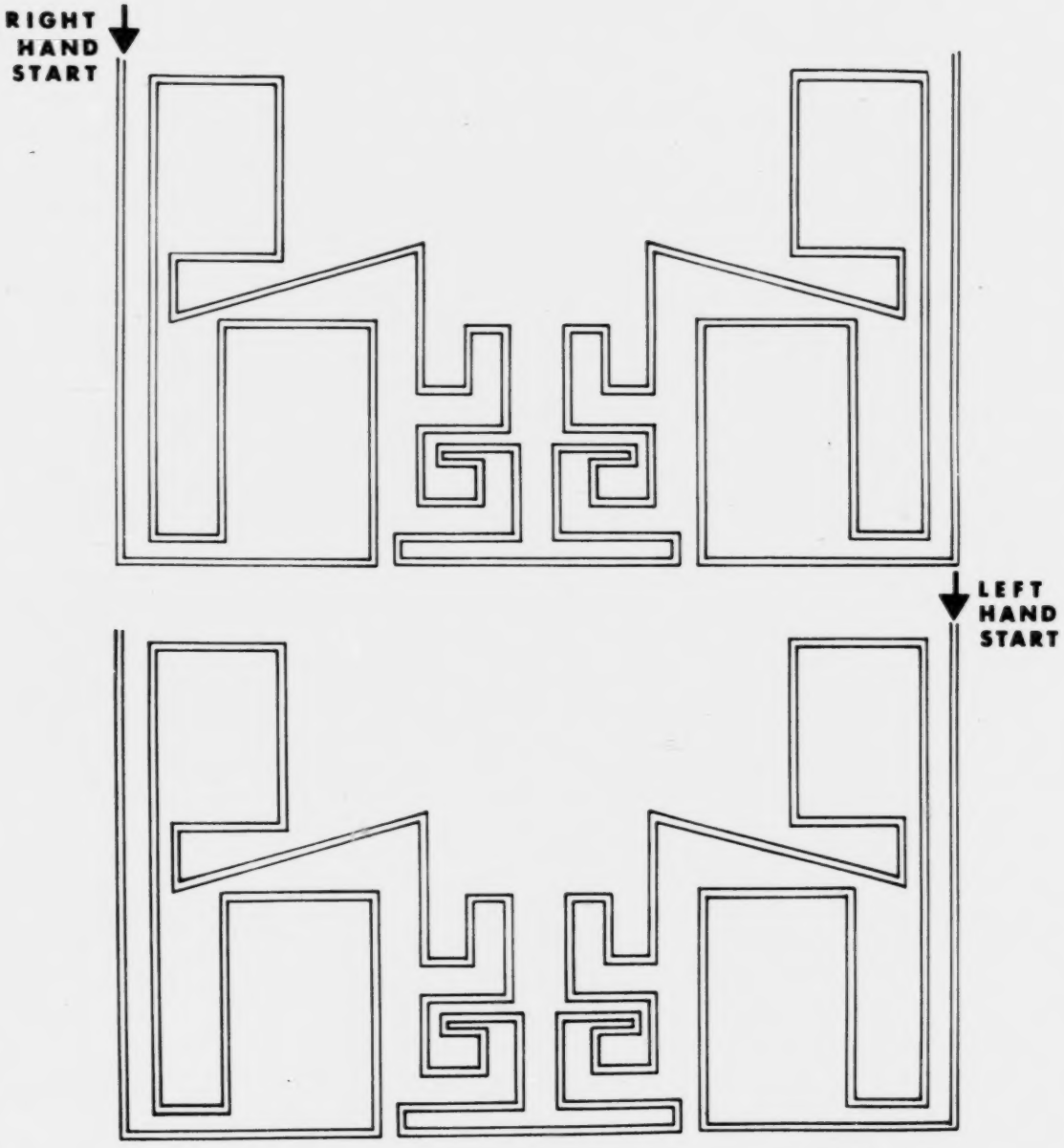


Fig. 2

Dichotic Task

Nine different dichotic tests were run, but data are reported here for only one: the consonant-vowel (CV) syllable stop consonant test. Six syllables, formed from the six English stops, /b, d, g, p, t, k/, followed by the vowel /æ/, were synthesized on the Haskins Laboratories parallel resonant synthesizer. A fully balanced, 60-item dichotic tape was then prepared. Each subject took this test twice in one day, with earphones reversed on the second run to distribute channel effects equally over the ears, and twice in a second day. This yielded a total of 240 trials per ear per subject.

Scoring

An adjusted difference score, right minus left (R-L), was computed for each subject, totaled over all runs on each task. For the handedness tasks, all scores, whether in seconds, number of completed items or kilograms of pull, were treated as frequency data. Using the normal distribution as an approximation to the binomial, the right hand score was expressed as a deviation from the expected mean, to yield a standard score ($Z = R\text{-}L/\sqrt{R+L}$).[1]

For the dichotic test, the phi-coefficient of correlation between performance and ear of presentation was computed. Kuhn (1972) has shown that this index compensates for variations in observed laterality effects due to variations in overall performance. Equivalent to R-L/R+L at 50% performance (where the possible ear difference is at a maximum), the coefficient systematically and symmetrically increases the weight attached to a given ear difference, as performance departs from this level, and so permits comparison among laterality effects independent of their associated levels of performance. It is therefore peculiarly apt for use in a study of individual differences.

RESULTS

We begin with results for the 22 unselected adults. Figure 3 (left side) presents histograms of individual scores on two handedness tests: tracing and scissors. Both tests yield a significant mean right-hand advantage, but the scatter of scores is wide, especially for the scissors task, and the distributions are negatively skewed. Other handedness tests showed a similar pattern.

---

[1]

$$Z = \frac{R - \frac{(R+L)}{2}}{\sqrt{(1/2)^2 N}}$$

(where R = right hand score
L = left hand score
N = R+L)

$$Z = \frac{R - \frac{(R+L)}{2}}{\sqrt{(R+L)/4}} = \frac{2R - (R+L)}{\sqrt{R+L}} = \frac{R-L}{\sqrt{R+L}}$$

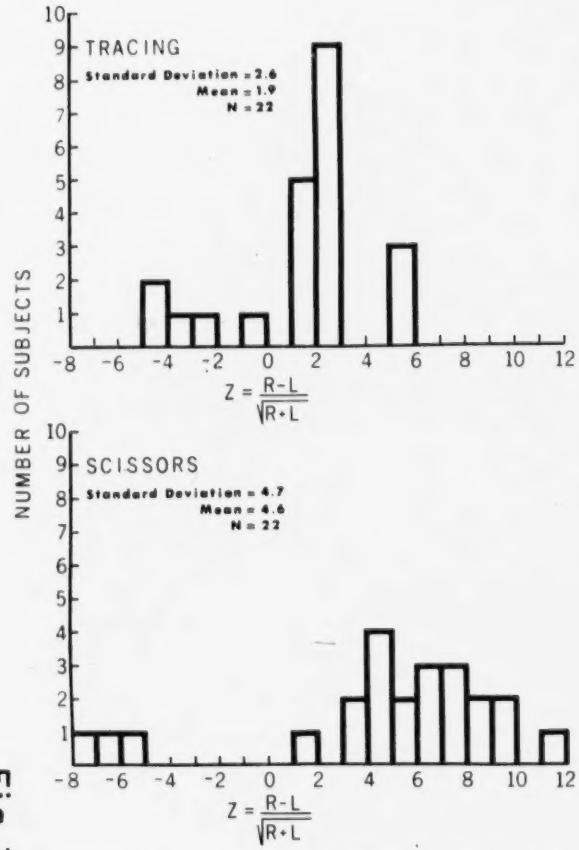
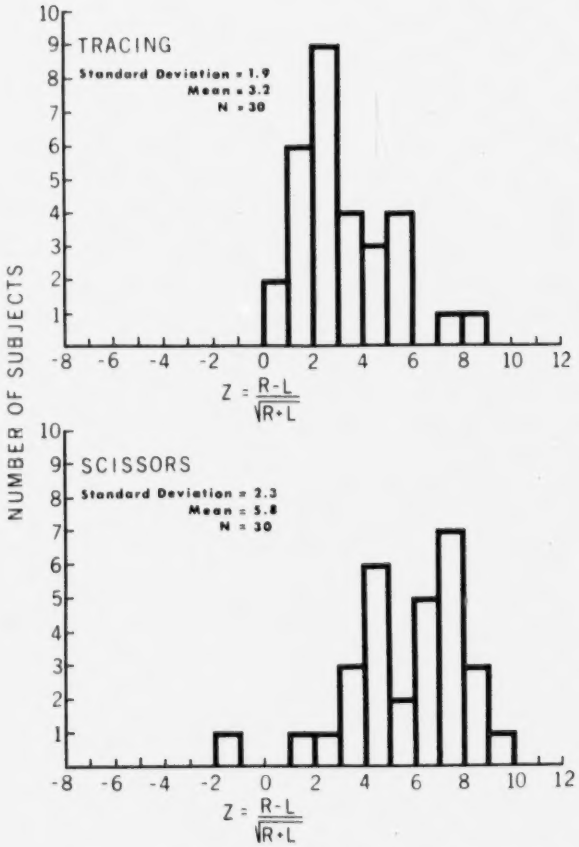


Fig. 3

Distribution of laterality indices on tracing and scissors tasks for unselected adults.



Distribution of laterality indices on tracing and scissors tasks for right-handed male adults.

Figure 3: Distribution of laterality indices on tracing and scissors tasks for 22 unselected adults (left) and 30 right-handed male adults (right).

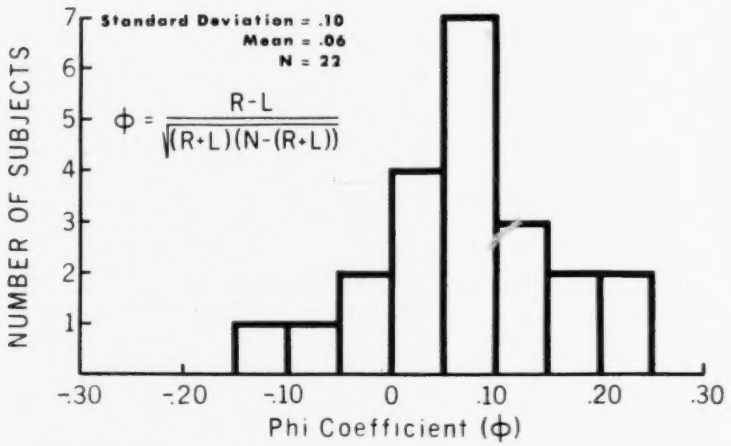
On the dichotic consonants test (Figure 4 top) the distribution is more or less symmetrical around a mean right-ear advantage, as measured by the phi-coefficient, of .06. Test-retest reliabilities for lateral differences on the several tasks. are moderately high (see Table 1), ranging from .69 for Crawford Screws to .93 for the scissors test. Table 2 displays intercorrelations among the tests. The lower four lines show values of the product moment correlation coefficient among the handedness tests: all are statistically significant and form, for this group of subjects, a relatively tight cluster. The top line shows values of the coefficient for the dichotic consonants test and each of the handedness tests: none of them reaches significance at the .05 level.

However, a composite index predicts the perceptual asymmetry considerably better than the single measures. Figure 5 plots normal deviates of the obtained ear advantage against normal deviates of the handedness tasks, weighted and combined according to the regression equation displayed on the figure. Four of the five handedness tasks--all except tapping--enter the equation and contribute significantly, at the .05 level or better, to the prediction. The multiple correlation coefficient is .69. The increase in the multiple coefficient over the simple coefficients suggests that the several handedness tasks measure distinct additive components of handedness.

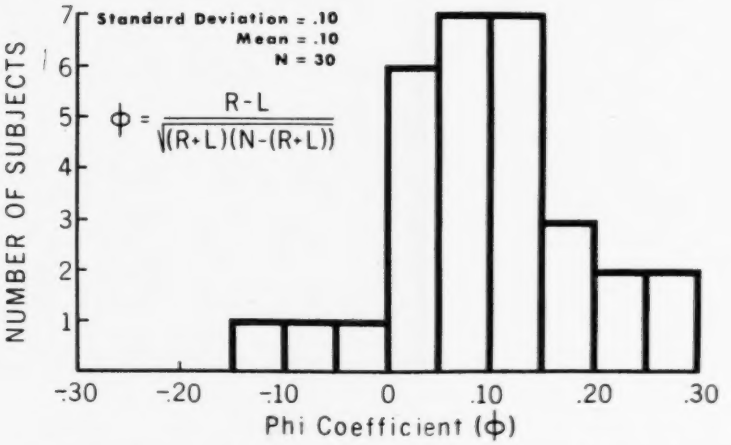
The data reported so far are perhaps open to the objection that the group of unselected adults included several left-handers for whom some relation between handedness and degree of cerebral dominance might be expected. A more telling test of the relation is provided by the results for the homogeneous group of 30 right-handed males.

Figure 3 (right side) displays their performance on the scissors and tracing tasks: the means for the right-handers are shifted to the right relative to those for the unselected subjects, and the variability, though still striking, has been reduced. Figure 4 (bottom) displays the distribution of ear advantages: the mean has again shifted to the right, but the standard deviation is unchanged. Table 3 displays the test-retest reliabilities for lateral differences: they range from .38 for the Crawford Pegs to .70 for the dichotic consonants (a value identical to that for the unselected group). These are surprisingly low, and it seems likely that more extensive testing is necessary for a relatively homogeneous group such as this if adequate reliabilities are to be reached. This conclusion is supported by the intercorrelations of Table 4. There we see that only one pair of handedness tasks (Crawford Pegs and Crawford Screws) shows any significant correlation. On the other hand, two tasks (Scissors, Tracing) show moderate, but significant correlations with the dichotic scores.

Finally, Figure 6 plots the multiple regression equation. Here, only two of the handedness tasks (Scissors, Tracing) contribute significantly, at the .05 level or better, to prediction of the ear advantage. As might be expected on statistical grounds, the reduced handedness range yields a lower correlation coefficient than was found for the unselected group of adults: .54 instead of .69. However, since the sample size was larger, the coefficient is significant at a higher level.



Distribution of laterality indices on dichotic consonants test for unselected adults.



Distribution of laterality indices on dichotic consonants test for right-handed male adults.

Fig. 4

Figure 4: Distribution of laterality indices on dichotic consonants test for 22 unselected adults (top) and 30 right-handed male adults (bottom).

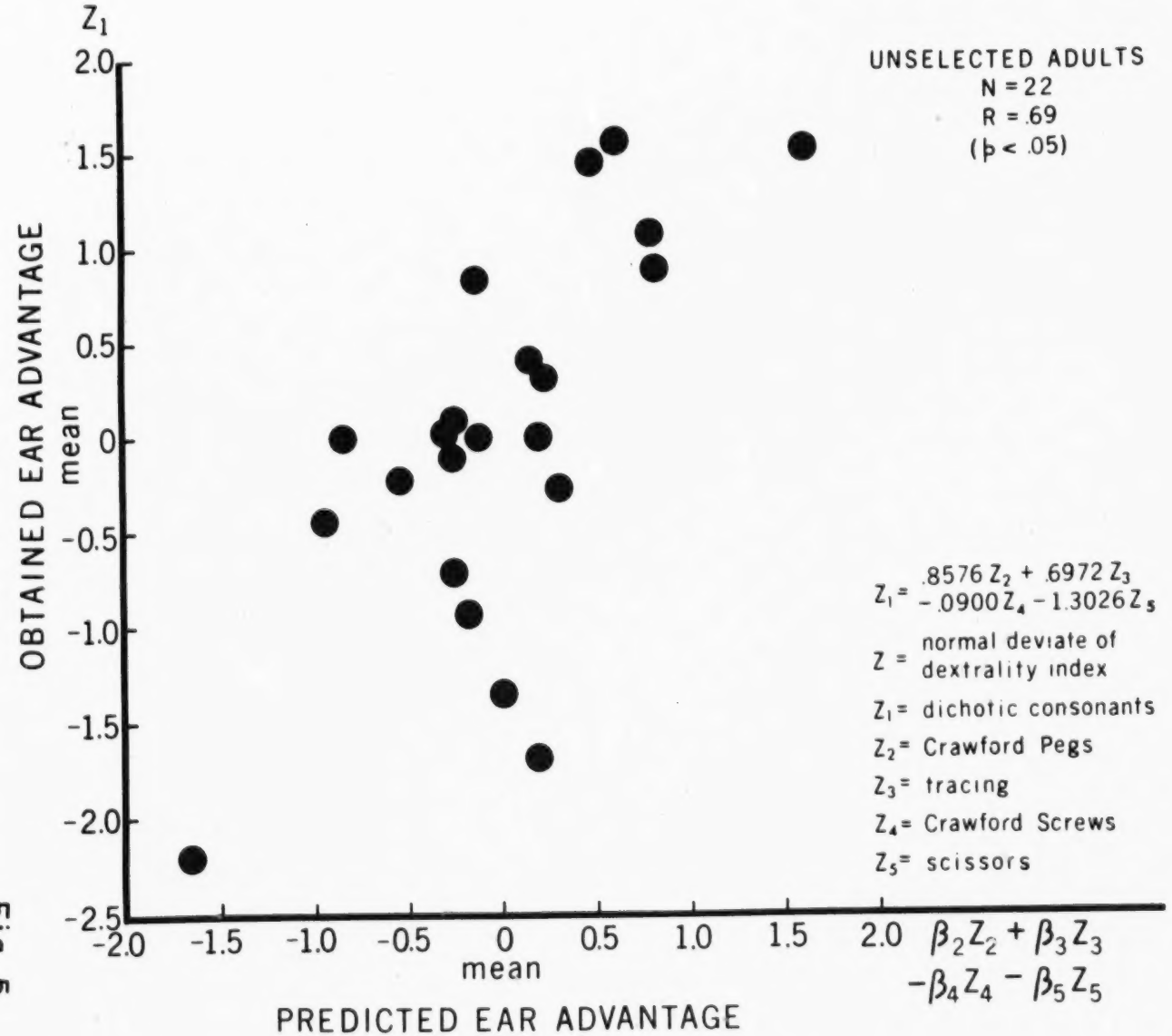


Fig. 5

Figure 5: Normal deviates of the obtained ear advantages as a function of normal deviates of four handedness tasks, weighted and combined according to the regression equation of 22 unselected adults.

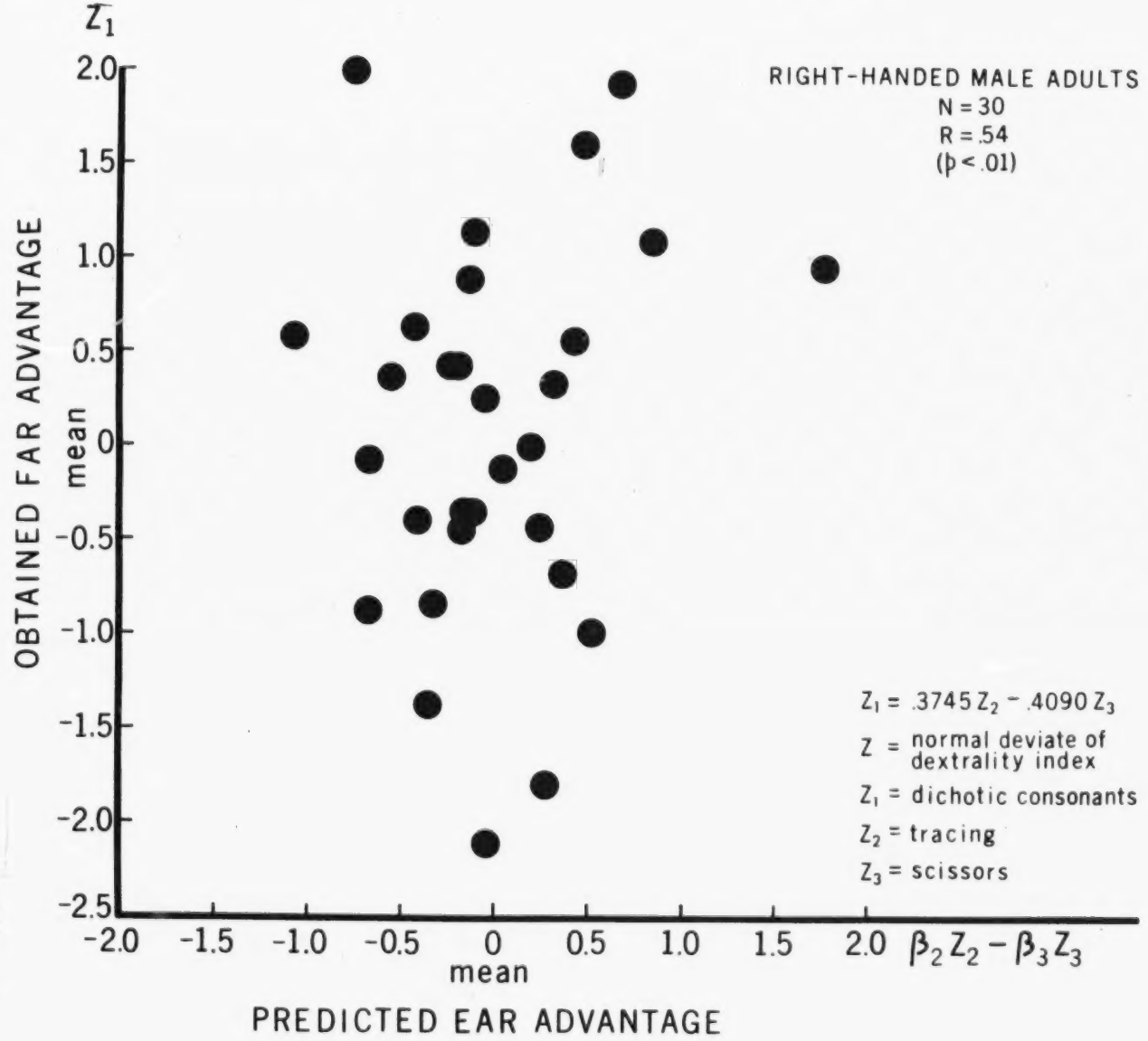


Fig. 6

Figure 6: Normal deviates of the obtained ear advantages as a function of normal deviates of two handedness tasks, weighted and combined according to the regression equation for 30 right-handed male adults.

TABLE 1

TEST-RETEST RELIABILITIES FOR UNSELECTED ADULTS (N = 22)

TEST	PEARSON R
DICHOTIC (CONSONANT)	.70
TAPPING	.80
SCISSORS	.93
TRACING	.80
CRAWFORD PEGS	.78
CRAWFORD SCREWS	.69

TABLE 2

INTERCORRELATIONS (PEARSON R) FOR PHI-COEFFICIENT OF DICHOTIC CONSONANT TEST AND Z-SCORES OF HANDEDNESS TESTS FOR UNSELECTED ADULTS (N = 22)

	Tapping	Scissors	Tracing	Cr. Pegs	Cr. Screws
Dichotic	-.05	-.14	.21	.21	.26
Tapping		.73***	.66***	.66***	.60**
Scissors			.76***	.81***	.53*
Tracing				.66***	.69***
Crawford Pegs					.66***

* $p < .05$
** $p < .01$
*** $p < .001$

TABLE 3

Test-retest reliabilities for right-handed male adults (N = 30)

Test	Pearson r
Dichotic (consonant)	.70
Tapping	.44
Scissors	.44
Tracing	.68
Crawford Pegs	.38
Crawford Screws	.44

TABLE 4

INTERCORRELATIONS (PEARSON R) FOR PHI-COEFFICIENT OF DICHOTIC CONSONANT TEST AND Z-SCORES OF HANDEDNESS TESTS FOR RIGHT-HANDED MALE ADULTS (N = 30)

	TAPPING	SCISSORS	TRACING	CR. PEGS	CR. SCREWS
DICHOTIC	-.17	.40*	.36*	.03	-.14
TAPPING		-.01	.05	.15	-.01
SCISSORS			-.03	.03	-.13
TRACING				.01	-.10
CRAWFORD PEGS					.35*

* $p < .05$

DISCUSSION

The results are consistent with the findings of Orlando (1971). Individual differences in the size of the ear advantage covary significantly with differences in the degree of measured handedness. Taken together, the two studies provide substantial support for the hypothesis that cerebral dominance for speech perception should be viewed as a continuum across individuals. The studies are, of course, restricted to a single type of perceptual process; phonetic recognition of English stop consonants. However, we may be justified in speculating on the implications of such findings, should they be confirmed and extended, for a model of the mechanism of cerebral dominance.

A main implication is that the concept of dominance, whether for language or handedness, must be expanded. As it stands, the concept is merely a summary restatement of the effects of unilateral lesions. Obviously, this cannot account for variations over more than three values: left, right, and center. If we take any group for whom this value is fixed for both language and handedness--say, a group of right-handers with left hemisphere specialized for the major language functions--we must now account for two facts. First, the scores of individuals within this group vary continuously on measures of both handedness and language function. Second, these two forms of continuous variation are correlated; that is to say, a significant proportion of the variance on both types of test has a common source.

The traditional concept of cerebral dominance (or hemispheric specialization) could, at best, account only for an association between handedness and speech. For example, if it could be shown that both speech and manual skills have a common source in, say, neural specialization for rapid, sequential behavior and, further, that there were good reasons why this capacity was concentrated in a single cerebral hemisphere, we would not be obliged to extend the concept of dominance beyond its present anatomical content. Just such an argument has, in fact, been made by Semmes (1968). From an extensive study of brain-injured war veterans she argues that "the phylogenetic trend toward increased localization of function" (cf. Geschwind, 1971; Geschwind and Levitsky, 1968) has issued in focal organization of the left hemisphere and its consequent specialization for "behaviors which demand fine, sensori-motor control, such as manual skills and speech" (p. 11).

However, the first fact--if it be one--namely, that lateralization for certain functions varies continuously across individuals [and, incidentally, perhaps within individuals across functions (cf. Day and Vigorito, 1972; Cutting, 1972)] cannot be accounted for without extending our concept of lateralization to include a dynamic, variable component. It seems, in fact, that we should be viewing lateralization not simply as a fixed anatomical characteristic, but rather as a process or function governing the relations between hemispheres, and open to variation within and across individuals. Just how to characterize this process we have, as yet, little knowledge to suggest.

Finally, we should stress that the data of the present study are no more than preliminary, both methodologically and substantively. Methodologically, future work will have to concentrate on selecting and refining the measures of both handedness and language dominance to achieve a fuller sampling of skills and higher test-retest reliabilities. Substantively, the study has examined

but one aspect of a single language function. Dichotic methods are necessarily restricted to the study of perceptual and short-term memorial processes. But within these limits, the technique has already been adapted to the study of a wide range of linguistic functions, from prosody to syntax and meaning.

Ultimately dichotic testing may even play a valuable clinical role, answering, in some measure, the need expressed by Luria when he wrote: "It is easy to see that our lack of knowledge concerning the degree of dominance of the hemisphere in different persons and with respect to different functions is a great handicap in the clinical investigation of patients with local brain lesions" (1966:90).

REFERENCES

Annett, M. (1970) A classification of hand preference by association analysis. Brit. J. Psychol. 61, 303-321.

Benton, A. L. (1965) The problem of cerebral dominance. Canad. Psychologist 6a, 332-348.

Benton, A. L., R. Meyers, and G. J. Polder. (1962) Some aspects of handedness. Psychiat. Neurol. 144, 321-337.

Bryden, M. P. (1965) Tachistoscopic recognition, handedness and cerebral dominance. Neuropsychologia 3, 1-8.

Bryden, M. P. (1970) Laterality effects in dichotic listening: Relations with handedness and reading ability in children. Neuropsychologia 8, 443-450.

Crawford, J. E. and D. M. Crawford. (1956) Small Parts Dexterity Test. (New York: Psychological Corporation).

Curry, F. K. W. (1967) A comparison of left-handed and right-handed subjects on verbal and non-verbal dichotic listening tasks. Cortex 3, 343-352.

Cutting, J. E. (1972) A parallel between encodedness and the magnitude of the right ear effect. Haskins Laboratories Status Report on Speech Research SR-29/30, 61-68.

Day, R. S. and J. M. Vigorito. (1972) A parallel between degree of encodedness and the ear advantage: Evidence from a temporal order judgment task. Paper read before 84th meeting of Acoustical Society of America, Miami Beach, Fla., November.

Geschwind, N. (1971) Some differences between human and other primate brains. In Cognitive Processes of Non-Human Primates, ed. by L. E. Jarrard. (New York: Academic Press).

Geschwind, N. and W. Levitsky. (1968) Human brain: Left-right asymmetries in temporal speech region. Science 161, 186-187.

Hecaen, H. and J. Ajuriaguerra. (1964) Les Gauchers. (Paris: Presses Universitaires de France).

Kimura, D. (1961a) Some effects of temporal-lobe damage on auditory perception. Canad. J. Psychol. 15, 156-165.

Kimura, D. (1961b) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.

Kuhn, G. (1972) The phi-coefficient as an index of ear differences in dichotic listening. Haskins Laboratories Status Report on Speech Research SR-29/30, 75-82.

Luria, A. R. (1966) Higher Cortical Functions in Man, tr. by Basil Haigh. (New York: Basic Books).

Milner, B., C. Branch, and T. Rasmussen. (1966) Evidence for bilateral speech representation in some non-right handers. Trans. Amer. Neurol. Assoc. 91A.

Orlando, C. (1971) Relationships between language laterality and handedness in eight and ten year old boys. Unpublished Ph.D. dissertation, University of Connecticut, Storrs.

Satz, P., K. Achenbach, and E. Fennell. (1967) Correlations between assessed manual laterality and predicted speech laterality in a normal population. Neuropsychologia 5, 295-310.

Satz, P., K. Achenbach, E. Pattishall, and E. Fennell. (1965) Order of report, ear asymmetry and handedness in dichotic listening. Cortex 1, 377-396.

Semmes, J. (1968) Hemispheric specialization: A possible clue to mechanism. Neuropsychologia 6, 11-26.

Subirana, H. (1958) The prognosis in aphasia in relation to the factor of cerebral dominance and handedness. Brain 8, 415-425.

Zangwill, O. (1960) Cerebral Dominance in Relation to Psychological Function. (Edinburgh: Oliver Boyd).

Zurif, E. B. and M. P. Bryden. (1969) Familial handedness and left-right differences in auditory and visual perception. Neuropsychologia 7, 179-187.

A Parallel Between Degree of Encodedness and the Ear Advantage: Evidence from a Temporal Order Judgment Task*

Ruth S. Day[+] and James M. Vigorito[+]
Haskins Laboratories, New Haven

Speech sounds are not speech sounds are not speech sounds. That is, some speech sounds appear to be more highly "encoded" than others (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). Let us illustrate this notion of encodedness in a very simple way.

Suppose we have a syllable consisting of a stop consonant followed by a vowel. We now break this syllable into two portions. The first segment contains all the information preceding the steady-state portion of the vowel. When we play it in isolation, several times in succession, listeners usually identify it as a coffee pot gurgle, a Model T sputter, or some other nonspeech sound. However, when we play the second portion of the split syllable, listeners have no difficulty in identifying it as the vowel /i/.

The point of this tape splicing demonstration is illustrated in Figure 1. The basic sound units of speech, the phonemes, are not added up like so many beads on a string, as shown on the left side of the display. Instead, there is an overlapping of linguistic segments as shown on the right side of the display. As these segments overlap, they undergo restructuring at the acoustic level.

Some speech sounds undergo more restructuring than others, as shown by the "split /p/" demonstration. The /p/ has a particular acoustic structure, namely one that is appropriate to be in initial position and followed by the vowel /i/. Therefore it cannot be recovered perceptually after the tape splicing procedure. Meanwhile the /i/ has undergone relatively little change as a function of context. Hence it can easily be recovered perceptually despite the fact that it has been spliced out of context.

Those speech sounds that undergo the most restructuring in the sound stream are said to be highly encoded, whereas those that undergo relatively little change as a function of neighboring phonemes are said to be less encoded. In general, stop consonants such as /p/ are highly encoded whereas vowels are relatively unencoded.

---

*Paper presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., December 1972.

[+]Also Department of Psychology, Yale University, New Haven.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

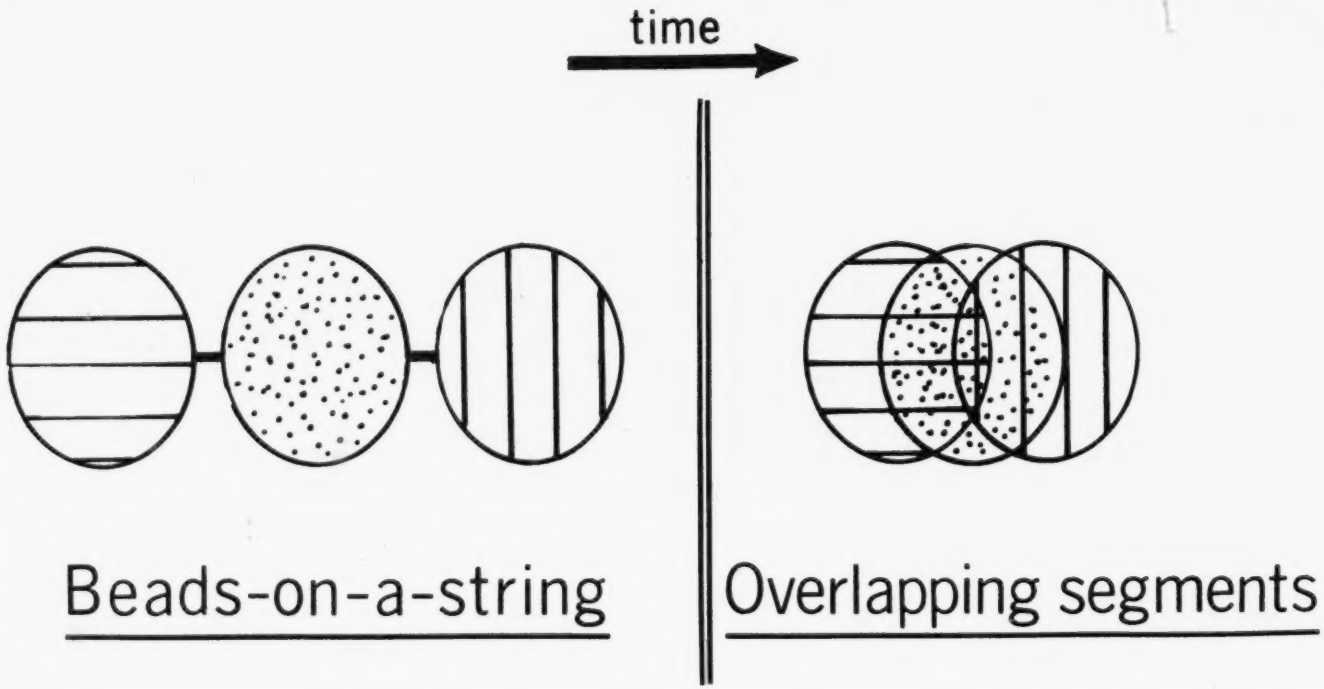


Fig. 1

Figure 1: Two views of the arrangement of phonetic units in time.

A summary of tape splicing results is shown in Figure 2. We have already considered stop-vowel syllables such as /pi/. What happens when we apply the same tape splicing procedure to other classes of speech sounds? Consider a fricative-vowel syllable such as /si/ as shown on the right side of the display. When listeners are asked to identify the first segment, they report hearing /s/ most of the time. This finding suggests that /s/ is not as highly encoded as the stop consonants. Nevertheless, it is not recovered perceptually as often as the vowel; therefore fricatives appear to be more highly encoded than vowels.

It may well be that there is an encodedness continuum for classes of speech sounds, with stop consonants at one end, vowels at the other end, and remaining classes such as fricatives and liquids falling in the middle.

At this point a word of caution is in order. There are several different ways to locate classes of speech sounds on such an encodedness continuum. Tape splicing experiments have been discussed here since they can be presented quickly and clearly. Other types of supporting evidence can also be presented, for example categorical perception for consonants versus vowels (Studdert-Kennedy, Liberman, Harris, and Cooper, 1970).

What happens when we pit phonemes within a given class against each other in a dichotic listening task? A right-ear advantage in dichotic listening is thought to reflect the participation of a special speech processing mechanism. Do the highly encoded speech sounds engage this mechanism to a greater extent than the less encoded sounds? Indeed, can we predict the magnitude of the ear advantage by placing classes of speech sounds along an encodedness continuum determined on independent grounds? Other investigators have shown that dichotic stimuli differing only in the initial stop consonant yield highly reliable right-ear advantages. Vowel contrast pairs, however, yield inconsistent results, with some studies obtaining a small ear advantage and others no ear advantage at all. Such results suggest that vowels can be perceived as speech or as nonspeech. For a recent review of this literature, see Studdert-Kennedy and Shankweiler (1970).

The present experiment compared the ear advantages of stops, liquids, and vowels. Note that we are interested in a <u>rank ordering</u> of these stimulus classes in terms of the ear advantage: stops should have the largest right-ear advantage, liquids less of a right-ear advantage, and vowels the least right-ear advantage.

METHOD

On each dichotic trial one of the items began 50 msec before the other. The subject's task was to determine which syllable began first. Thus he had to make a temporal order judgment (TOJ). There were three tests which differed only in their vocabulary: the stop test used /bæ, dæ, gæ/, the liquid test /ræ, læ, wæ/, and the vowel test /i, a, u/. All stimuli were prepared on the parallel resonance synthesizer at the Haskins Laboratories, then arranged into dichotic tapes using the pulse code modulation system. The syllables were highly identifiable, as determined by a binaural pretest.

The 16 subjects received all three dichotic tests. The listeners were right-handed, native English speakers, and had no history of hearing trouble.

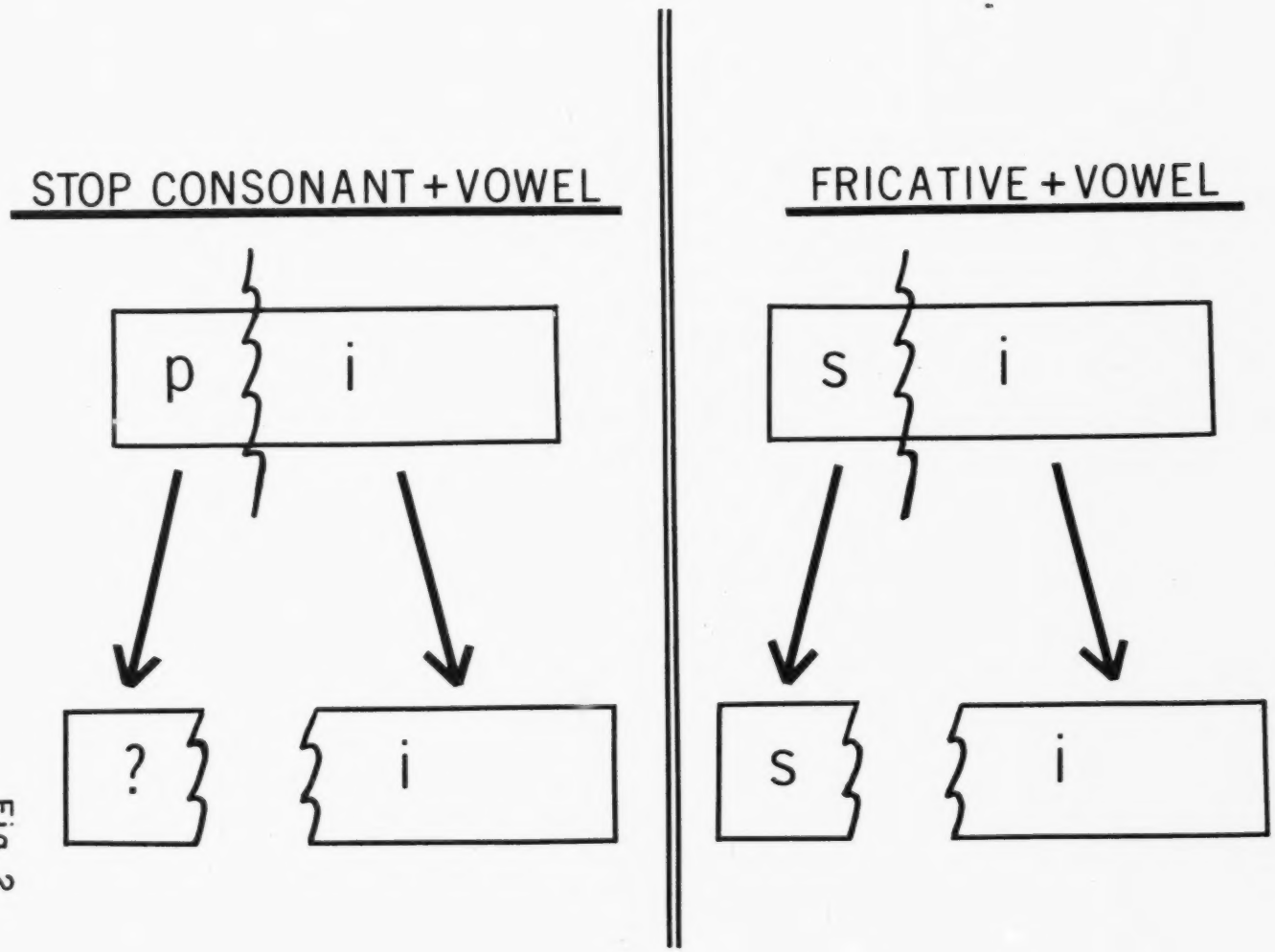


Fig. 2

Figure 2: Summary of tape splicing experiments.

Responses were scored in the following way. Given that a particular ear received the leading stimulus, what percent of the time was the subject correct in determining that that item did indeed lead? Ear difference scores were then obtained by subtracting percent correct TOJ for one ear from percent correct TOJ for the other ear.

RESULTS

The results are shown in Figure 3. The horizontal line indicates no ear advantage; here performance for the two ears is comparable. The region above this line indicates a right-ear advantage; that is, the right-ear score surpassed the left-ear score by the percent shown on the ordinate. The region below the horizontal line indicates a left-ear advantage in a comparable fashion.

Stops yielded a right-ear advantage, liquids a slight right-ear advantage, and vowels a left-ear advantage. Thus, as we moved along the encodedness continuum from stops to liquids to vowels, the right-ear advantage became reduced and finally disappeared. This is exactly the rank ordering predicted.

DISCUSSION

There appears to be a parallel between encodedness and the ear advantage. If indeed the right-ear advantage in dichotic listening reflects the operation of a special processing mechanism, then the present data suggest that the highly encoded speech sounds require the services of this mechanism to a greater extent than do the less encoded sounds.

The present data are compatible with those of Cutting (1972) who used consonant-consonant-vowel (CCV) syllables. The first consonant was always a stop consonant (/g/ or /k/), the second was a liquid (/l/ or /r/), and the following vowel was either /æ/ or /ɛ/, yielding eight syllables in all. A different syllable was presented to each ear at the same time and, for a given block of trials, the subject was asked to monitor one ear and report only the syllable presented to it. Ear advantage scores were obtained by subtracting percent correct for one ear from the percent correct for the other ear; this analysis was performed separately for each phoneme class. Again the rank ordering of phonemes in terms of the right-ear advantage was stops > liquids > vowels. However, the liquids yielded a sizable right-ear advantage while the vowels yielded no ear advantage in Cutting's experiment. Thus the data in the present TOJ experiment have in a sense been shifted "leftward" by comparison. There are several possible reasons for this shift. Cutting used CCV syllables; hence the speech processor may well have been engaged early in stimulus presentation, such that the liquids were clearly perceived in a "speech mode" while vowels were so perceived roughly half the time. Subjects in the present experiment heard liquid-vowel syllables and vowels in isolation; hence there were no stop consonants to engage the speech mechanism before the target contrasts were presented. Another possibly relevant factor is the type of dichotic tasks used. Cutting's ear-monitoring task required only identification of a syllable and not a judgment about relative timing. It could be that judgments about relative onset time are best handled in the nonspeech hemisphere, which would tend to depress the overall level of right-ear advantage scores. Finally, the different ear advantage levels obtained in the two experiments may have occurred simply because different subjects were used. Individuals differ in the extent

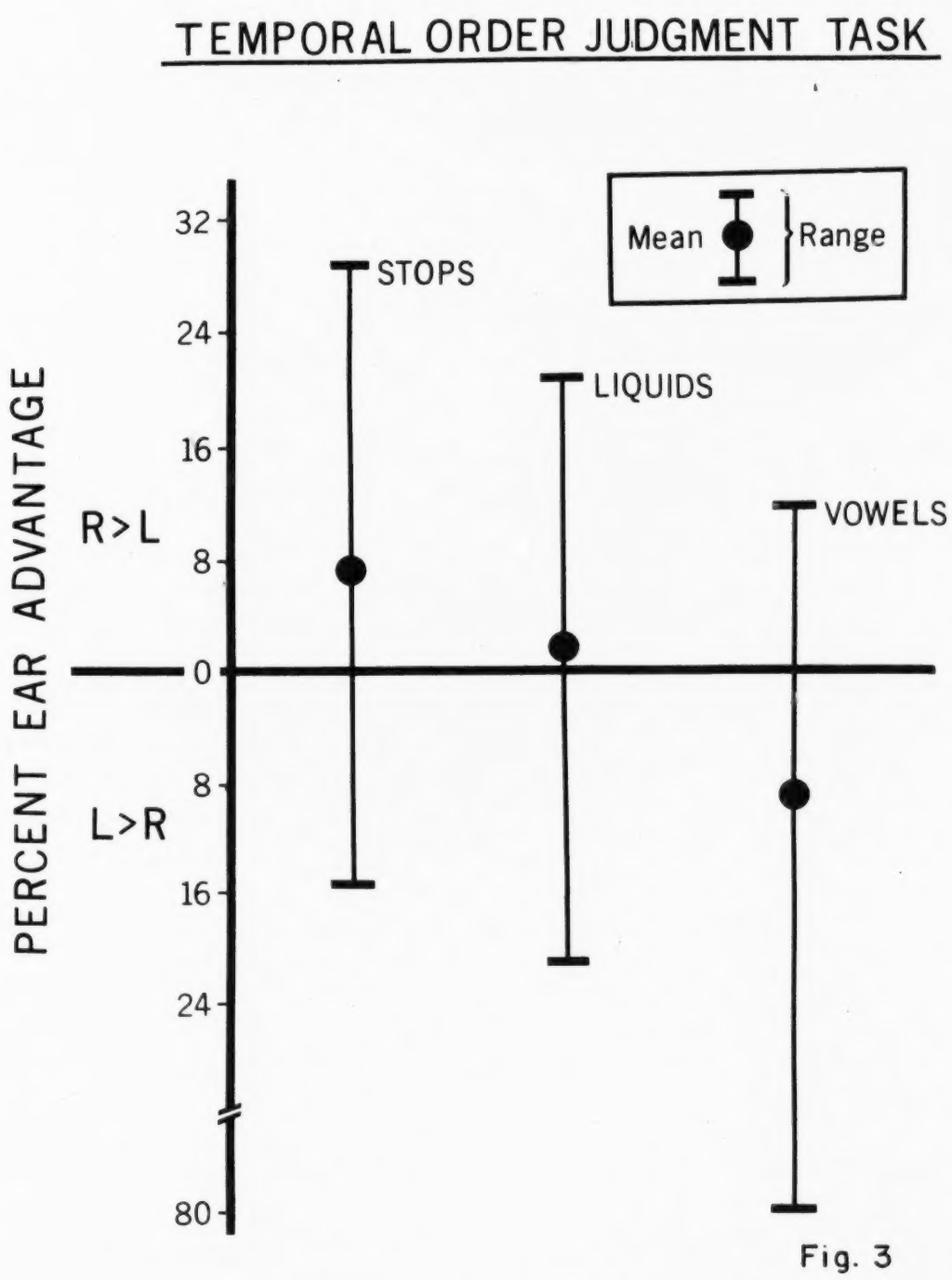


Figure 3: Ear advantages for stops, liquids, and vowels.

to which they show a right-ear advantage for dichotic speech items. Nevertheless, both experiments lend support to the notion that there is a parallel between the encodedness of speech sounds and the ear advantage.

REFERENCES

Cutting, J. E. (1972) A parallel between degree of encodedness and the ear advantage: Evidence from an ear-monitoring task. Paper presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla. (Also in Haskins Laboratories Status Report on Speech Research SR-29/30, 61-68, as: A parallel between encodedness and the magnitude of the right ear effect.)

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Memory for Dichotic Pairs: Disruption of Ear Report Performance by the Speech-Nonspeech Distinction*

Ruth S. Day,[+] James C. Bartlett,[+] and James E. Cutting[+]
Haskins Laboratories, New Haven

When several dichotic pairs are presented in rapid succession, the best strategy is usually to segregate the items by ear of arrival. Subjects spontaneously adopt this strategy in the free recall situation. They report all the items presented to one ear before reporting those presented to the other ear. This finding was first reported by Broadbent (1954) and has been replicated many times. Forced order of report experiments also demonstrate the effectiveness of the ear report technique. Subjects are instructed to use a particular report method for a given block of trials. For example, in the time-of-arrival method, subjects must report both items in the first pair, then those of the second, and finally those of the third. Recall accuracy for the ear-of-arrival method is superior to that of the time-of-arrival method in the ordered recall situation (Broadbent, 1957; Moray, 1960).

The ear report effect for rapid dichotic pairs is a hardy one. In addition to occurring in both free recall and ordered recall situations, it occurs in a wide variety of circumstances, including word lists and digit lists (Bryden, 1964), and lists of different lengths (Bryden, 1962).

Ear report performance is reduced only when there are sufficiently strong cues favoring another mode of organization. One such cue is stimulus class.[1] In a free recall task, Gray and Wedderburn (1960, Experiment II) used both words and digits and alternated them between the ears over successive dichotic pairs. For example, one ear received "MICE-5-CHEESE" while the other ear received "3-EAT-4." Subjects reported the items by stimulus class ("MICE-EAT-CHEESE, 3-5-4") more often than by ear. Comparable results have been obtained for recall accuracy in the ordered report situation using mixed dichotic pairs of unrelated words and digits (Yntema and Trask, 1963).

---

*Paper presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., 1 December 1972.

[+]Also Department of Psychology, Yale University, New Haven.

[1]Another cue is presentation rate. At slow rates, subjects perform best when they report the items in pair-wise fashion, by time of arrival (Bryden, 1962). In the present paper we are concerned only with rapid rates of presentation (about 2 pairs/sec).

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

Not all stimulus class distinctions are equally effective in reducing the ear report effect. In one experiment, the individual syllables of a three-syllable word were paired with digits, for example, "EX-2-PATE" and "6-TIR-9" (Gray and Wedderburn, 1960, Experiment I). Report by class, namely "EX-TIR-PATE, 6-2-9," occurred only when subjects were told that the syllables form an English word; uninformed subjects never reported a whole word. Hence the cohesiveness of items within a given class may be an important factor in overriding the ear report effect.

The present study was designed to assess the effect of a different type of stimulus class distinction: speech vs. nonspeech. To what extent can this distinction override the normally useful ear report method?

METHOD

The speech stimuli were the natural speech syllables /ba, da, ga/. The nonspeech stimuli were 500, 700, and 1000 Hz tones. All stimuli were 300 msec in duration and the intensity envelopes of the nonspeech stimuli were made to resemble those of the speech stimuli. Stimulus editing and dichotic tape preparation were accomplished using the pulse code modulation system at Haskins Laboratories.

There were two basic types of trials, as shown in Figure 1. On segregated trials, all three items of a given class went to the same ear: all the speech items went to the left ear and all the nonspeech to the right ear, or vice versa. On cross-over trials, the items of a given class switched between the ears. In the 2-1 case, two items of the same class were presented to a given ear, followed by one item from the other class. In the 1-2 case, a single item from a class was presented to a given ear, followed by two items from the other class. Finally, in the 1-1-1 case, the class of item changed with each successive pair for a given ear.

Note that for all trials, every dichotic pair consisted of speech to one ear and nonspeech to the other. Furthermore, the same six items were presented in every triplet; only the sequence (in time) and location (with respect to ear) of these items distinguished the triplets. The various types of trials were randomly mixed on the same tape. The interval between successive pairs was 250 msec, while the interval between successive triplets was 6 sec.

A partial report procedure was used. Before a block of triplets began, the subject was asked to report the order in which a given subset occurred. He was asked to report the order of a given stimulus class on some blocks; thus, he reported only the order of the three speech stimuli or the order of the three nonspeech stimuli (labeled "Hi, Med, Lo"). For other blocks, he was asked to report by ear; thus, he reported the order of the three left-ear stimuli or that of the three right-ear stimuli. This procedure substantially reduced memory load. All responses were written with the letters B, D, G and H, M, L serving to designate speech and nonspeech stimuli, respectively.

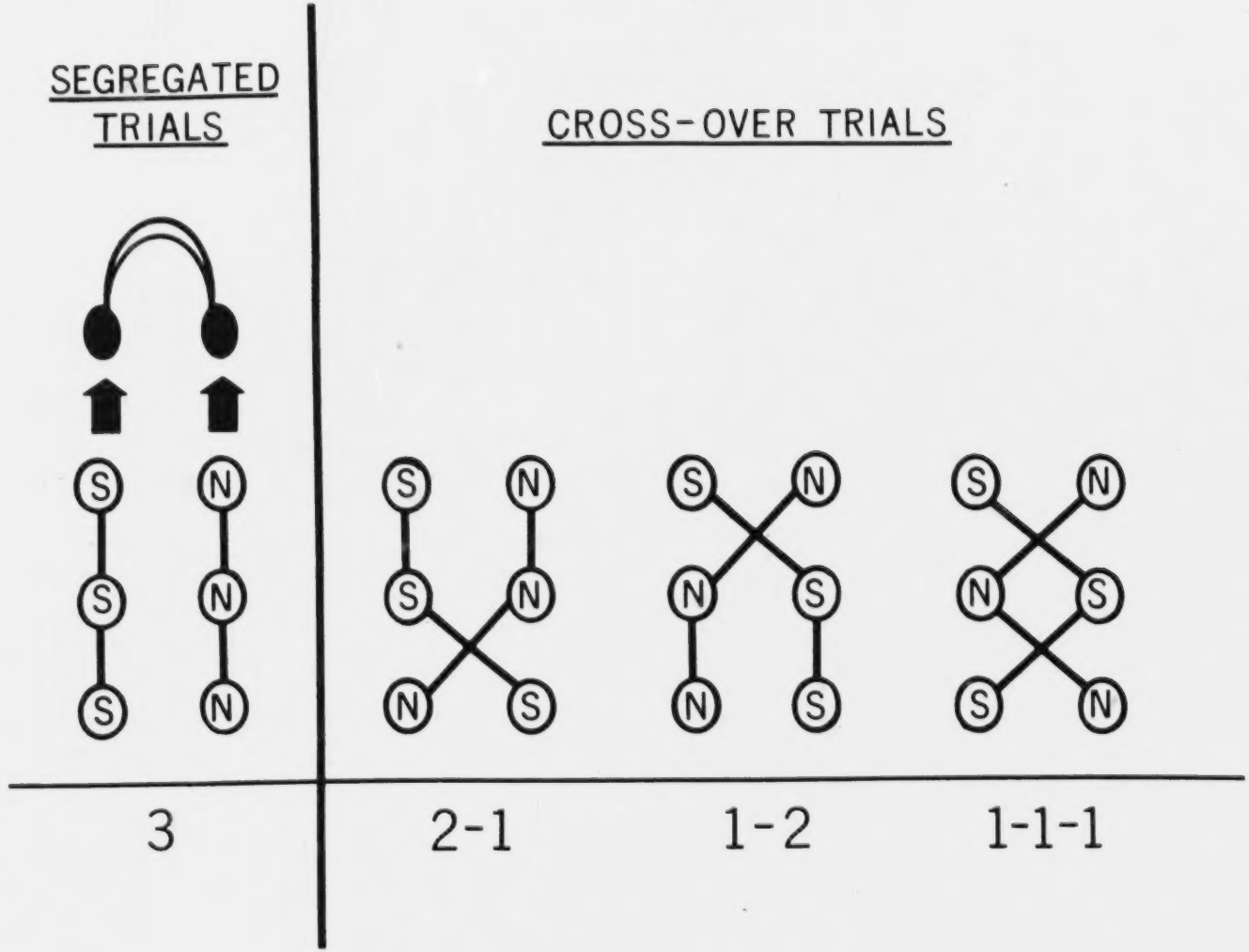


Fig. 1

Figure 1: Types of trials for speech/nonspeech dichotic triplets.

RESULTS

The major results are shown in Figure 2. On segregated trials, performance was excellent, whether subjects had to report by ear or by stimulus class. On cross-over trials, report by class remained an excellent report technique, but report by ear suffered greatly. Subjects were unable to monitor a single ear successfully when it received both speech and nonspeech stimuli. Stimulus class has an almost "magnetic attraction" in these cross-over trials. If one listens to them with no particular strategy in mind, the items of a given class seem to pull one's attention around, first to one side of the head, later to the other, depending on the type of trial presented.

As we have seen, the stimulus class distinction between speech and nonspeech is so powerful that subjects have difficulty in reporting by ear, even when specifically asked to do so, and when only three items are involved. When we break these data down into more detail, as shown in Figure 3, we see that report by both speech and nonspeech remained high over all types of trials. Although performance on nonspeech was a few percentage points above that for speech, this difference was not statistically significant. The ear report data show a clear contrast between segregated and cross-over trials, and also a difference among the various kinds of cross-over trials. Performance on cross-over trials was best for ear report when two stimuli of the same class were presented to a given ear before a switch was made to the other class; performance was worst here when stimulus class changed with every pair. In terms of the ear scores themselves, there were no significant differences in performance levels between the two ears.

DISCUSSION

This experiment shows that speech vs. nonspeech is a very powerful distinction--powerful enough to reduce the effectiveness of the normally useful ear report method in dichotic memory tests. Elsewhere we have shown that the speech-nonspeech distinction is powerful enough to change the outcomes of other well-known paradigms (Day and Cutting, 1971a). For example, consider identification performance on single pair dichotic trials. When both stimuli are speech (S/S), errors occur. It is as if the two speech stimuli are sent to a single processing system. This system cannot handle two items at once, and hence, errors occur. When both items are nonspeech (NS/NS), errors also occur, again suggesting that both nonspeech stimuli are overloading a single processing system. Since different ear advantage results occur in these two cases, a right-ear advantage for S/S (Kimura, 1961) and a left-ear advantage for NS/NS (Kimura, 1964), it appears that different processing systems are being called into service. Recently we showed that when one item is speech and the other is nonspeech (S/NS), no errors occur (Day and Cutting, 1971a). Subjects are equally able to determine which speech and nonspeech items are presented, even when the items have been drawn from a test vocabulary that is as large as four speech stimuli and four nonspeech stimuli (Day and Cutting, 1971b). It appears then that the speech and nonspeech items of a single S/NS pair are sent to different processing systems, each of which can perform its work without interference from the other.

If indeed there are separable processing systems for speech and nonspeech, then this notion would help explain the present data. Stimuli from the two classes are sent to different perceptual systems, making it difficult to reorganize them in another way, say, by ear-of-arrival. It is not clear when the speech-nonspeech distinction is made--during ongoing perception or later, during response organization. The present data suggest that the distinction may be made quite early.

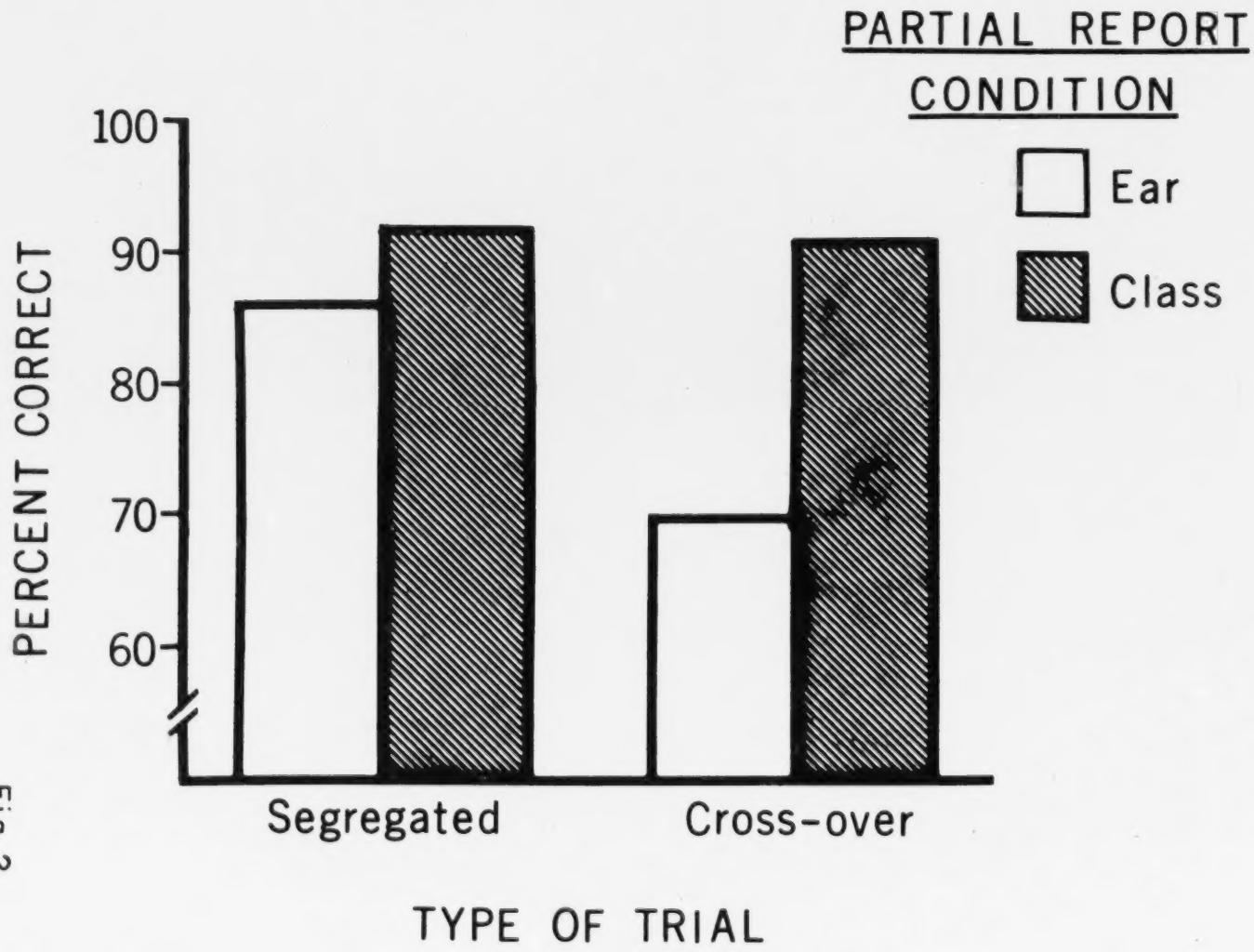


Fig. 2

Figure 2: Percent correct partial report by ear and class on segregated and cross-over trials.

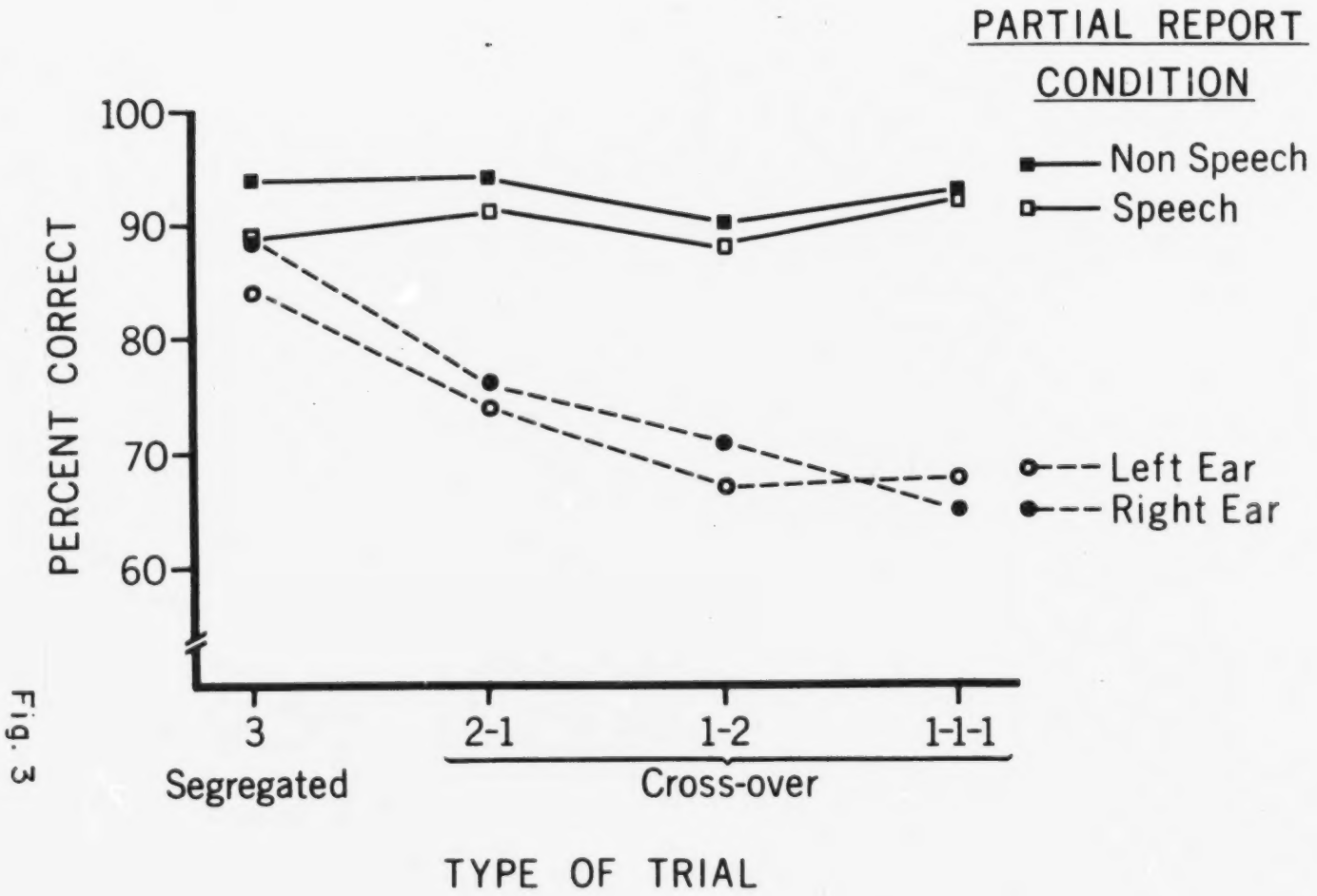


Fig. 3

Figure 3: Percent correct partial report for all conditions on all types of trials.

Experiments that inform the subject which subset of stimuli to report, either before or after a given triplet is presented, are in progress and hopefully will help determine when the speech-nonspeech distinction is made.

In any event, it is clear that the speech-nonspeech distinction is a very fundamental one, one that is powerful enough to change the results of some standard paradigms. By varying the types of stimuli used in the present paradigm, we may be able to learn more about what makes speech, "speech" and nonspeech, "nonspeech."

REFERENCES

Broadbent, D. E. (1954) The role of auditory localization and attention in memory span. J. Exp. Psychol. 47, 191-196.

Broadbent, D. E. (1957) Immediate memory and simultaneous stimuli. Quart. J. Exp. Psychol. 9, 1-11.

Bryden, M. P. (1962) Order of report in dichotic listening. Canad. J. Psychol. 16, 291-299.

Bryden, M. P. (1964) The manipulation of strategies of report in dichotic listening. Canad. J. Psychol. 18, 126-138.

Day, R. S. and J. E. Cutting. (1971a) Perceptual competition between speech and nonspeech. J. Acoust. Soc. Amer. 49, 85(A). (Also in Haskins Laboratories Status Report on Speech Research SR-24, 35-46.)

Day, R. S. and J. E. Cutting. (1971b) What constitutes perceptual competition in dichotic listening? Paper presented at the Eastern Psychological Association meeting, New York, April.

Gray, J. A. and A. A. Wedderburn. (1960) Grouping strategies with simultaneous stimuli. Quart. J. Exp. Psychol. 12, 180-184.

Kimura, D. (1961) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exp. Psychol. 16, 355-358.

Moray, N. (1960) Broadbent's filter theory: Postulate H and the problem of of switching time. Quart. J. Exp. Psychol. 12, 214-220.

Yntema, D. B. and F. P. Trask. (1963) Recall as a search process. J. Verb. Learn. Verb. Behav. 2, 65-74.

Ear Advantage for Stops and Liquids in Initial and Final Position*

James E. Cutting+
Haskins Laboratories, New Haven

Some parts of the sound pattern of speech appear to require more linguistic processing than others. In general, consonants require more of this special processing than vowels, and some consonants require more than others. One measure of the amount of linguistic processing required for the different classes of speech sounds may be found in the results of dichotic listening tasks.

When one speech stimulus is presented to one ear, and a similar speech stimulus to the other ear at the same time, the subject often has difficulty reporting both of them correctly. Typically, he is able to report the speech stimulus presented to the right ear better than the one presented to the left. There is a two-fold explanation for this right-ear advantage. First, most people process speech primarily in one hemisphere of the brain--usually the left. Second, the auditory pathway from the right ear to the left hemisphere is dominant over the pathway from the left ear to the left hemisphere. Thus, the speech stimulus presented to the right ear has direct access to the speech processor, whereas the left ear stimulus appears to travel a more circuitous route to the processor by way of the right hemisphere and the corpus callosum (for a review, see Studdert-Kennedy and Shankweiler, 1970).

One measure of the amount of linguistic processing required for different classes of speech sounds appears to be the magnitude of the right-ear advantage in dichotic listening tasks. Some speech sounds yield large right-ear advantages, while others yield little or no advantage for either ear. Recently, there is evidence that speech sounds array themselves on a continuum of right-ear advantages. Cutting (1972a) used a dichotic ear-monitoring task and found a large right-ear advantage for stop consonants, a reduced right-ear advantage for liquids, and no ear advantage for vowels. A similar pattern of results has been found for stops, liquids, and vowels in a dichotic temporal order judgment task (Day and Vigorito, 1972).

One explanation for this ear advantage continuum involves the notion of "encodedness." Liberman, Cooper, Studdert-Kennedy, and Shankweiler (1967) have defined encodedness as the general amount of acoustic restructuring a phoneme

---

*This is a longer version of a paper submitted for presentation at the 85th Convention of the Acoustical Society of America, Boston, April 1973.

+Also Yale University, New Haven.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

undergoes in different contexts. Some phonemes undergo a great deal of acoustic change, while others undergo less change. The highly encoded phonemes, stop consonants, yield the largest right-ear advantages, while the relatively unencoded vowels generally yield no ear advantage. Liquids, which are less encoded than stops but more encoded than vowels, yield intermediate results. Day (in press) and Cutting (1972a) have suggested that there are two continua, an "encodedness" continuum and an ear advantage continuum, which are functionally parallel.

The present study was designed to study the effect of syllable position on the ear advantage in dichotic listening for certain phoneme classes. If the only variable responsible for the magnitude of the ear advantage is encodedness, then variations in other parameters, such as syllable position, should have no effect. To insure that encodedness was held constant stimuli were synthesized so that the acoustic structure of initial and final consonants were mirror images of each other.

GENERAL METHOD

Stimuli. Twenty-four syllables were prepared on the Haskins Laboratories parallel resonance synthesizer: twelve were consonant-vowel (CV) syllables and twelve were vowel-consonant (VC) syllables. All possible combinations of the consonants /b, g, l, r/ and the vowels /i, ae, ɔ/ were used: thus, there were six stop-vowel syllables (/bi, bae, bɔ, gi, gae, gɔ/), six vowel-stop syllables (/ib, aeb, ɔb, ig, aeg, ɔg/), six liquid-vowel syllables (/li, lae, lɔ, ri, rae, rɔ/), and six vowel-liquid syllables (/il, ael, ɔl, ir, aer, ɔr/). The stimuli were 325 msec in duration and all had the same falling pitch contour. Except for pitch, the CV and VC syllables were exact mirror images of one another. As shown in Figure 1, the acoustic structure of /bae/ is identical to that of /aeb/ except that the time axis has been reversed. The syllables /lae/ and /ael/ show the same relationship, as do all other CV-VC pairs with the same consonant and vowel. This reversal was accomplished using a revised version of the pulse code modulation (PCM) system at Haskins Laboratories (Cooper and Mattingly, 1969). This system enables the experimenter to reverse the time axis of a stimulus in the memory buffer, flipping it end to end, without changing any other parameters. The command for this operation is FLIP. Stop consonants which preceded the same vowel differed only in the direction and extent of the second formant transition. Liquids which preceded the same vowel differed only in the direction and extent of the third formant transition. This same pattern is true for stops and liquids which followed the same vowel.

Subjects. Sixteen Yale undergraduates served as subjects in two tasks. They were all right-handed native American English speakers with no history of hearing difficulty. Subjects were tested in groups of four, with stimuli played on an Ampex AG500 tape recorder and sent through a listening station to Grason-Stadler earphones.

TASK I: IDENTIFICATION

A brief identification test was run to assess the quality of the stimuli.

Tapes and procedure. Subjects listened to one token of each stimulus to familiarize themselves with synthetic speech. They then listened to two binaural

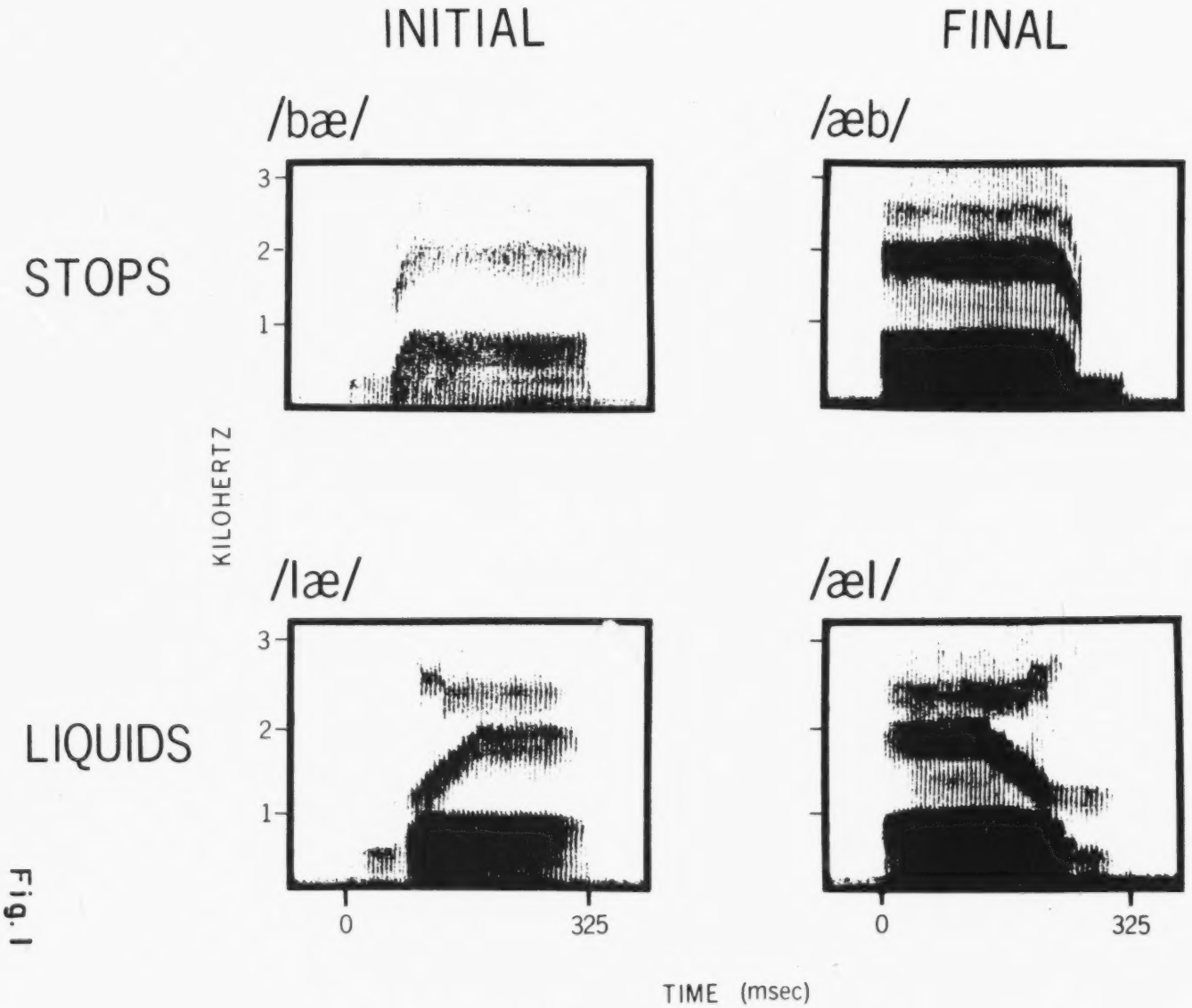


Fig. 1

Figure 1: Sound spectrograms of /b/ and /l/ in initial and final syllable position. Note that initial and final consonants are mirror images of each other.

identification tapes, each with 120 items. One tape consisted of the stop stimuli and the other consisted of the liquid stimuli. Each of the twelve stimuli within the stop and the liquid sets was presented ten times in random sequence with a two-second interstimulus interval. Subjects were asked to identify only the consonant in each of the stimuli, writing B or G for the stops, and L or R for the liquids.

Results. The stimuli were highly identifiable. All were identified at a rate of nearly 90 percent.

TASK II: EAR MONITORING

Tapes and procedure. The same 24 stimuli were used; however, this time, instead of presenting one stimulus at a time, two stimuli were presented simultaneously, one to each ear. These dichotic tapes were prepared on the PCM system. Two tapes were made for the set of stop stimuli and two tapes for the set of liquids. Within each set two rules governed the pairing of stimuli: (1) both stimuli in a dichotic pair were either CV syllables or VC syllables, and (2) the two stimuli shared neither the same consonant nor the same vowel. Thus, /bae/ was paired with /gi/ and /gɔ/, while /aeb/ was paired with /ig/ and /ɔg/. The same pattern was followed for the liquid stimuli. The reason for using different vowels in the dichotic pairs became evident in a pilot study. Liquid stimulus pairs such as /li/-/ri/ were perceived as a single ambiguous stimulus: subjects heard only one item and it appeared to be an acoustic average of the two stimuli. Cutting (1972b) has described this type of psychoacoustic fusion as a low-level, perhaps peripheral, process. One way to eliminate this type of fusion is to pair stimuli that differ in vowels. This procedure allows more central processing to occur.

Each tape consisted of 72 dichotic pairs: (6 possible pairs within a syllable class) X (2 syllable classes, CV and VC) X (2 channel arrangements per pair) X (3 replications). Two such tapes with different random orders were prepared for the stop stimuli. Two similar tapes were prepared for the liquids. All tapes had a three-second interval between pairs. Subjects listened to two passes through each 72-item tape for both stops and liquids, yielding a total of 576 trials per subject.

Subjects were instructed to monitor only one ear at a time, and to write down the consonant that was presented to that ear, B or G, L or R. For each set of stimuli the order of ear monitoring was done in the following manner: half the subjects attended first to the right ear for a quarter of the trials, then to the left ear for half the trials, and then back to the right ear for the last quarter (RLLR). The other half of the subjects attended in the opposite order (LRRL). There was a brief rest between blocks of 72 trials. The order of headphone assignments and the order of listening to the stop and liquid dichotic tapes were also counterbalanced across subjects.

Results. The task was quite difficult: overall performance for all stimulus pairs was 67 percent. There was no difference in the overall performance for the stop stimuli and the liquid stimuli. Syllable position, however, proved to be important: performance on the initial consonants was slightly better than on the final consonants. Subjects were 70 percent correct for both initial stops and initial liquids, while they were only 64 percent correct for final stops and final liquids. This net 6 percent difference was significant, $F(1,15) = 15.4$, $p < .005$.

Both initial and final stops yielded a right-ear advantage. Furthermore, both ear advantages were of the same general magnitude. While monitoring initial stops, subjects were 73.8 percent correct for the right ear and 67.4 percent correct for the left ear. Subtracting left-ear scores from right-ear scores, they had a net 6.4 percent right-ear advantage. Final stops yielded a similar pattern. Subjects were 66.8 percent correct for the right ear and 60.9 percent for the left ear, yielding a net 5.9 percent right-ear advantage. Both right-ear advantages were significant, $F(1,15) = 8.5$, $p < .025$, and there was no significant difference between them. Figure 2 shows the net ear advantages for both types of stop pairs.

The liquids yielded a different pattern of results than the stops. Initial liquids yielded a right-ear advantage, but final liquids did not. While monitoring initial liquids, subjects were 72.9 percent correct for the right ear and only 66.1 percent correct for the left ear, yielding a net 6.8 percent right-ear advantage. This right-ear advantage, like those of the initial and final stops, was significant. Final liquids, however, yielded no significant ear advantage. Subjects were 62.9 percent correct for the right ear and 64.5 percent correct for the left ear, yielding a net 1.6 percent left-ear advantage. Thus, unlike the stops, the liquids show a pattern of results which varies according to the position of the target phoneme within the syllable. The (Ear) X (Syllable class) interaction for the liquid stimuli was significant, $F(1,15) = 6.4$, $p < .025$. Figure 2 shows the net ear advantages for the liquid pairs.

DISCUSSION

Degree of encodedness is a measure of the simultaneous transmission of speech sounds. The more highly encoded the speech sound the more its acoustic structure is folded into the acoustic structure of neighboring sounds. Differing degrees of encodedness appear to require differing degrees of special linguistic processing, and this processing appears to require a facility unique to the left hemisphere. Furthermore, encodedness of classes of speech sounds and the resulting ear advantage that those sounds yield in a dichotic listening task appear to be directly correlated. Phonemes which are highly encoded typically yield large right-ear advantages, phonemes which are relatively unencoded typically yield small or no ear advantages, and phonemes which are intermediate on the encodedness continuum typically yield intermediate ear advantages. If no linguistic principle other than encodedness influences the direction and magnitude of the ear advantage, then one would expect that, when the acoustic structure of the phonemes is held constant, ear advantages would be comparable for those phonemes in both CV and VC syllables. Indeed, in the present study stop consonants showed this relationship: comparable right-ear advantages were found for initial and final stops. Liquids, however, did not show this relationship. Initial liquids yielded a right-ear advantage, whereas final liquids did not, even though both types of liquids were made to have the same acoustic structure (but reversed in time).

Stop consonants are more highly encoded than liquids. Two recent studies have shown that stop consonants yield larger right-ear advantages than liquids (Cutting, 1972a; Day and Vigorito, 1972). The results of the present study show the same general pattern. Collapsing over the syllable position in which the stops and liquids occur, stops yielded a 6 percent right-ear advantage and

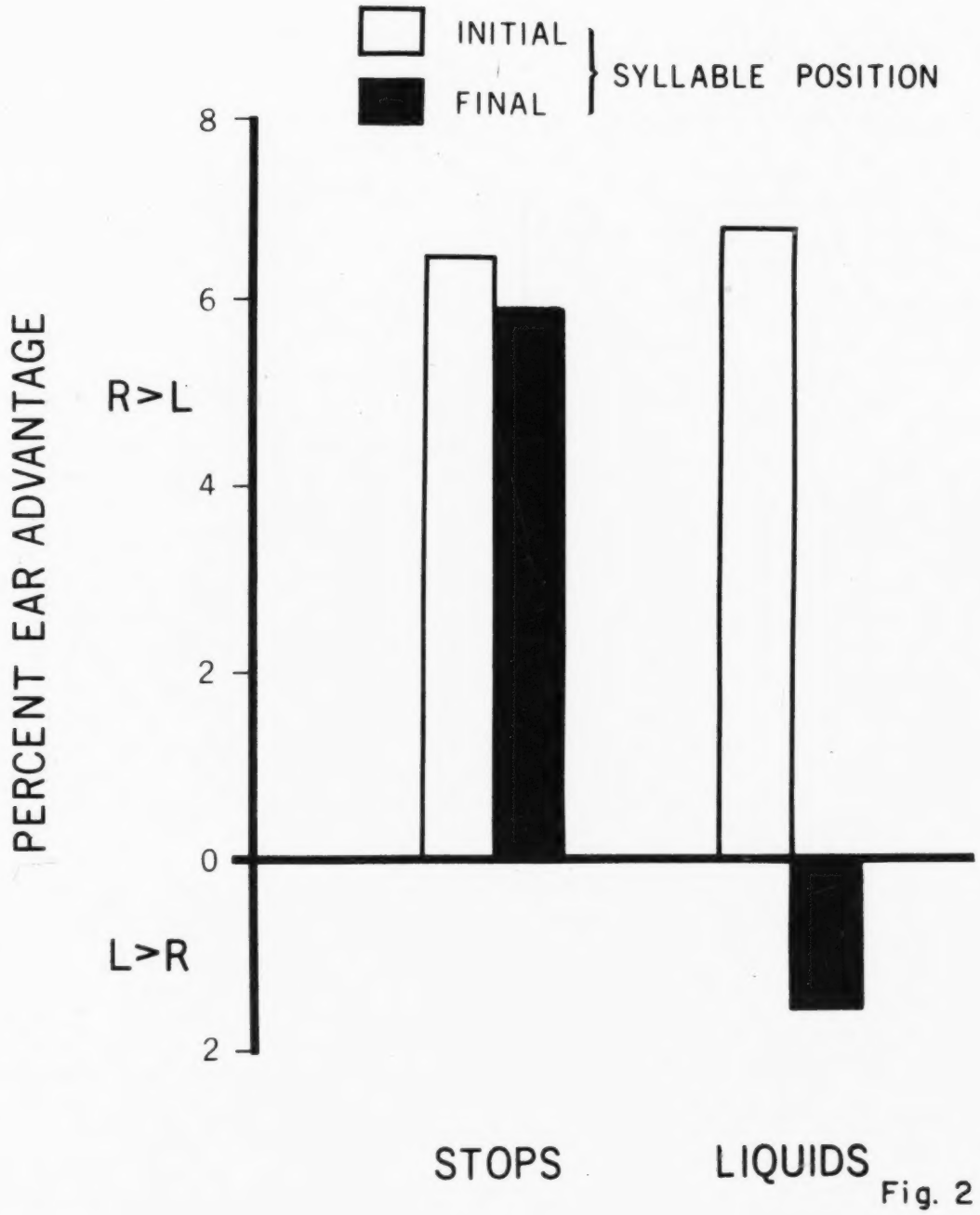


Figure 2: Mean ear differences in the ear-monitoring task.

liquids yielded a 3 percent right-ear advantage. However, when we look only at the ear advantages for initial stops and initial liquids, we find that both yield right-ear advantages of approximately 6 percent at comparable performance levels. Perhaps the most satisfactory explanation for this discrepancy may be found in the magnitude of the ear advantages. Note that the right-ear advantages in the present study are comparatively small in relation to those of previous studies. Let us compare the results of the present study with those of Cutting (1972a). Both were ear monitoring tasks, yet the results of the present study show ear advantages of 6 percent for stops and liquids in initial position, whereas those of the previous study show ear advantages of 12 and 9 percent for stops and liquids respectively.[1] Perhaps in the present study ear advantage scores were too small to manifest differences between stops and liquids in initial position.

Liquids as consonants and liquids as vowels. A right-ear advantage in dichotic listening is typically found for consonants, while no ear advantage is typically found for vowels. With this scheme in mind we might interpret the results of the present study in the following manner: initial liquids yield results which are typical of consonants, while final liquids yield results which are typical of vowels. Let us pursue this idea further. Liquids are maverick phonemes. In several distinctive feature systems (Jakobson, Fant, and Halle, 1951; Halle, 1964) they are considered to have both consonantal (consonant-like) and vocalic (vowel-like) features. Perhaps it is the initial liquids which are more like consonants and the final liquids which are more like vowels.

Consider differences in phonetic transcription: initial and final liquids are often represented by different phonetic symbols. For example, the two /r/-sounds in the word RAIDER may be transcribed differently. A typical phonetic transcription of RAIDER is /redɚ/. The initial /r/ is treated as a consonant, while the final /ɚ/ is treated as a vowel[2] (see Bronstein, 1960:116ff). A similar treatment may be found for /l/. Following Bronstein the two /l/-sounds in LADLE may be phonetically transcribed in a different manner. One such transcription is /ledɬ/. The initial /l/ is considered to be "light," whereas the final /ɬ/ is considered to be "dark" and is identified with a bar across the middle of the symbol. Bronstein considers the dark /ɬ/-sound to be similar to a back vowel. Perhaps it is the light /l/ which functions more like a consonant,

---

[1] A possible explanation of this decrease in the right-ear advantage may lie in the pairings of the stimuli: in the present study items in a dichotic pair differed in both consonant and vowel. The data of Studdert-Kennedy, Shankweiler, and Pisoni (1972) show that smaller right-ear advantages are obtained when consonant-target stimuli have different vowels than when they have the same vowel. Thus, the right-ear advantage for pairs like /bi/-/tu/ is smaller than for pairs like /bi/-/ti/. In the present study dichotic pairs were of the type /bi/-/gae/ and /li/-/rae/ rather than /bi/-/gi/ and /li/-/ri/.

[2] Brackets [---] are often used to signify phonetic transcriptions while slashes /---/ are often used to signify phonemic transcriptions (Chomsky and Halle, 1968:65). I have adopted to convention of using slashes for both for the sake of simplicity.

and the dark /ɫ/ which is more like a vowel. No phonemes other than /l/ and /r/ are transcribed differently as a function of their position within a syllable.

Initial and final liquids also show different developmental patterns. All allophones of /r/ and /l/ appear to be difficult for the young speaker to produce. Templin (1966) has noted that the two liquids, along with /s/, are the most difficult phonemes to master. Upon re-examination of previous findings, however, we find that initial liquids are generally easier to produce than final liquids. The data of Curtis and Hardy (1959) show that children who have difficulty with /r/-sounds have much less trouble pronouncing the initial /r/ than final /r/. The data of Templin (1957) show the same relationship in normal children for the two types of /l/: initial /l/ is easy for children to produce, whereas they have great difficulty with final /l/. No phonemes other than /l/ and /r/ show this differential developmental pattern as a function of syllabic position.

Initial and final liquids can be distinguished on several bases, including phonetic transcriptions and developmental patterns. Perhaps these are among the reasons that liquids yield different ear advantages in dichotic listening. Except for the time reversal, the liquids in the present study are acoustically identical, yet perceptually they are not identical.

CONCLUSION

Encodedness appears to be a useful and highly accurate predictor of the direction and magnitude of the ear advantage for classes of speech sounds in dichotic listening tasks. It has been defined as the general amount of acoustic restructuring that a phoneme undergoes in various contexts. Highly encoded phonemes (stop consonants) typically yield the largest right-ear advantages; less highly encoded phonemes (liquids) typically yield smaller right-ear advantages; and relatively unencoded phonemes (vowels) typically yield no ear advantage. Encodedness, however, cannot account for the direction and magnitude of ear advantages for all phoneme classes in all situations. Other, second-order linguistic factors may produce differential ear advantages within a phoneme class, even when the acoustic structure (encodedness) of the phonemes is held constant. For liquids one such factor is syllable position. They yield ear advantages more like those of highly encoded consonants when they appear in initial position, and yield results more like those of unencoded vowels when they appear in final position.

REFERENCES

Bronstein, A. J. (1960) The Pronunciation of American English. (New York: Appleton-Century-Crofts).

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper and Row).

Cooper, F. S. and I. G. Mattingly. (1969) Computer-controlled PCM system for investigation of dichotic speech perception. J. Acoust. Soc. Amer. 46, 115(A).

Curtis, J. F. and J. C. Hardy. (1959) A phonetic study of the misarticulation of /r/. J. Speech Hearing Res. 2, 244-256.

Cutting, J. E. (1972a) A parallel between degree of encodedness and the ear advantage: Evidence from an ear-monitoring task. Paper presented at the 84th meeting of the Acoustical Society of America, Miami, 1972. (Also in Haskins Laboratories Status Report on Speech Research SR-29/30, 61-68, as: A parallel between encodedness and the magnitude of the right ear effect.)

Cutting, J. E. (1972b) A preliminary report on six fusions in auditory research. Haskins Laboratories Status Report on Speech Research SR-31/32 (this issue).

Day, R. S. (in press) Engaging and disengaging the speech processor. In Hemispheric Asymmetry of Function, ed. by Marcel Kinsbourne. (London: Tavistock).

Day, R. S. and J. M. Vigorito. (1972) A parallel between encodedness and the ear advantage: Evidence from a temporal-order judgment task. Paper presented at the 84th meeting of the Acoustical Society of America, Miami, 1972. [Also in Haskins Laboratories Status Report on Speech Research SR-31/32 (this issue).]

Halle, M. (1964) On the bases of phonology. In The Structure of Language, ed. by J. A. Fodor and J. J. Katz. (Englewood Cliffs, N. J.: Prentice Hall).

Jakobson, R., C. G. M. Fant, and M. Halle. (1951) Preliminaries to Speech Analysis. (Cambridge, Mass.: MIT Press).

Liberman, A. M., F. S. Cooper, M. Studdert-Kennedy, and D. Shankweiler. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Studdert-Kennedy, M., and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Studdert-Kennedy, M., D. Shankweiler, and D. B. Pisoni. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. Cog. Psychol. 3, 455-466.

Templin, M. (1957) Certain Language Skills in Children. Monograph # 26 (Minneapolis: University of Minnesota Press).

Templin, M. (1966) The study of articulation and language development during the early school years. In The Genesis of Language, ed. by F. Smith and G. A. Miller. (Cambridge, Mass: MIT Press).

A Right-Ear Advantage in the Retention of Words Presented Monaurally

M. T. Turvey,[+] David Pisoni,[++] and Joanne F. Croog[+]
Haskins Laboratories, New Haven

Subjects were presented lists of 16 words at the rate of one word every two seconds in a probe memory paradigm. A list and its probe were presented in a random basis to either the right ear or the left ear. A distinct right-ear superiority was found in both the primary and secondary memory components of the short-term retention function.

INTRODUCTION

Several lines of evidence suggest that speech perception is markedly different from nonspeech auditory perception and that special processors may be involved (see Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Studdert-Kennedy, in press; Studdert-Kennedy and Shankweiler, 1970). This view is supported, in part, by clinical and laboratory observations indicating that speech and nonspeech are processed in different parts of the brain. Thus, Kimura (1961) discovered with normal subjects that if pairs of contrasting digits were presented simultaneously to right and left ears, those received by the right ear were more accurately reported; while if contrasting melodies were simultaneously presented, those received by the left ear were more accurately reported (Kimura, 1964). The right-ear superiority for dichotically presented verbal items has been repeatedly confirmed for both meaningful and nonsense speech (e.g., Broadbent and Gregory, 1964; Bryden, 1963; Curry and Rutherford, 1967; Shankweiler and Studdert-Kennedy, 1967). The left-ear superiority in the recall of dichotically presented nonspeech sounds has been confirmed for musical sequences (Darwin, 1969), sonar signals (Chaney and Webster, 1965), environmental noises (Curry, 1967), and clicks (Murphy and Venables, 1970).

Kimura (1961) attributed the right-ear advantage for dichotically opposed speech signals to the predominance of the left hemisphere for speech perception (in the majority of individuals) and to the functional prepotency of the crossed over the uncrossed auditory pathways. Evidence for the latter notion has come from several sources (e.g., Bocca, Caleano, Cassinari, and Migliavacca, 1955; Milner, Taylor, and Sperry, 1968; Rosenzweig, 1951).

[+]Also University of Connecticut, Storrs.

[++]Also Indiana University, Bloomington.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

For the most part the available evidence suggests that ear asymmetry occurs only under conditions of dichotic opposition; ear asymmetry with monaural presentation has been rarely, if ever, observed (Kimura, 1967; Satz, 1968). There seems to be at least one good reason why the demonstration of asymmetry should be restricted to dichotic stimulation. Milner, Taylor, and Sperry (1968) observed that under dichotic stimulation right-handed, commissurectomized patients were able to report verbal stimuli presented to the right ear, but not those presented to the left ear. On the other hand, with monaural presentation they performed equally well with either ear. This has been interpreted to mean that the ipsilateral pathway from left ear to left hemisphere is suppressed during dichotic stimulation, and with the callosal pathway sectioned, stimulation reaching the right hemisphere by means of the contralateral path is unable to transfer to the left hemisphere.

These data of Milner, Taylor, and Sperry (1968) (and also, Sparks and Geschwind, 1968) and their interpretation justify the rationale for the laterality effect proposed by Kimura (1961, 1964). Consider the stiuation in which under dichotic stimulation, normal left-hemisphere dominant subjects correctly perceive a left-ear speech input. We may presume that the input has been suppressed ipsilaterally but that it has reached the right hemisphere by the contralateral path and from there it has transferred across the callosal route to the left hemisphere where it is processed. Thus verbal items presented simultaneously to the right and left ears may be viewed as converging on the left hemisphere by different routes: those from the right ear go by the direct contralateral routes while those from the left ear travel an indirect and somewhat longer path, crossing first to the right hemisphere then across the commissure to the left. We should suppose, therefore, that the left-ear input, having to take an indirect route to the speech processor, is at a distinct disadvantage. Hence, the ear asymmetry occurs under dichotic stimulation.

In the present experiment we sought to determine whether ear differences might be demonstrated monaurally if the speech processor and its attendant speech-memory systems were severely taxed. Notable exceptions to the "asymmetry only under dichotic stimulation" rule are experiments by Bakker (1967, 1969) and Bakker and Boeynga (1970) which suggest that the retention of word lists is superior for monaural right-ear presentation, with the magnitude of the effect varying with list length (somewhat unsystematically) and recall method. Our experiment makes use of the Waugh and Norman (1965) probe short-term memory (STM) paradigm. This paradigm has the advantage of allowing for the separation of the two hypothesized memory systems--short-term store (STS) or primary memory, and long-term store (LTS) or secondary memory--which purportedly underlie the retention of material over brief periods of time. Our original expectations were that if a monaural right-ear advantage occurred it would probably be restricted to the STS component because of the close relation between this store and perceptual processes. However, the evidence presented here will show that both STS and LTS benefit from right-ear presentation.

METHOD

Subjects

The subjects were 34 undergraduates from the University of Connecticut who participated in the experiment as part of a course requirement. All subjects

were native speakers of English and handedness was not a criterion for participation in the experiment.

Lists

Thirty-six lists of 16 unrelated words were constructed. Words were drawn from the Thorndike and Lorge (1944) count in a quasi-random way, such that the sums of Thorndike-Lorge word frequencies in each list differed by less than 10%.

A probe word was chosen for each list so that items in positions 3, 5, and 7 were tested two times each for right- and left-ear presentation while items in positions 11, 13, and 15 were tested four times each for right- and left-ear presentation.

Procedure

Lists were tape-recorded at a rate of 2 sec per word. Before the presentation of each list a bell signaled the subject to be alert.

Each subject was instructed to fix his attention on every word as it was presented and not to review old words. At the end of a list a bell sounded again, followed by the probe word. The subject then attempted to respond with that word in the list that followed the probe word.

Each subject was encouraged to guess and each was given 25 sec to write down his response before the next list began. Four unanalyzed practice lists preceded the 36 experimental lists. Three, 2-min rest periods were allowed.

Each list and its respective probe word were presented to either right or left ear via a set of Grason-Stadler TDH-39 earphones. The position of the probe word was also randomized from list to list so that before presentation of any given list the subject did not know on which ear the list would be heard or what item would be probed for. The earphones were alternated across subjects to balance for possible channel effects.

RESULTS

The proportion of correct responses for each position tested is shown in Figure 1. The total number of correct responses per subject was significantly higher on a Wilcoxon matched-pairs test for right-ear presentation than for left-ear presentation, $\underline{P} < .001$. Of the 34 subjects, 26 showed a raw score right-ear superiority while only three subjects gave a left-ear superiority; the remaining five subjects performed equally well with either ear.

It is generally argued that the retention of material over brief periods of time is jointly determined by STS and LTS (cf. Kintsch, 1970). On the view that the two storage systems are stochastically independent, Waugh and Norman (1965) suggested that in the probe paradigm the probability of recalling an item in position i, $\underline{P}(R_i)$ is: $\underline{P}(R_i) = \underline{P}(STS_i) + \underline{P}(LTS) - \underline{P}(STS_i)\,\underline{P}(LTS)$. Where $\underline{P}(STS_i)$ and $\underline{P}(LTS)$ are the probabilities of recalling an item from short-term and long-term storage respectively. It is assumed that $\underline{P}(LTS)$ is independent of an item's position in the list and that $\underline{P}(STS_i)$ is maximal for the most recent item, and decreases monotonically as a function of the distance from the end of the list, reaching zero after approximately seven to ten intervening

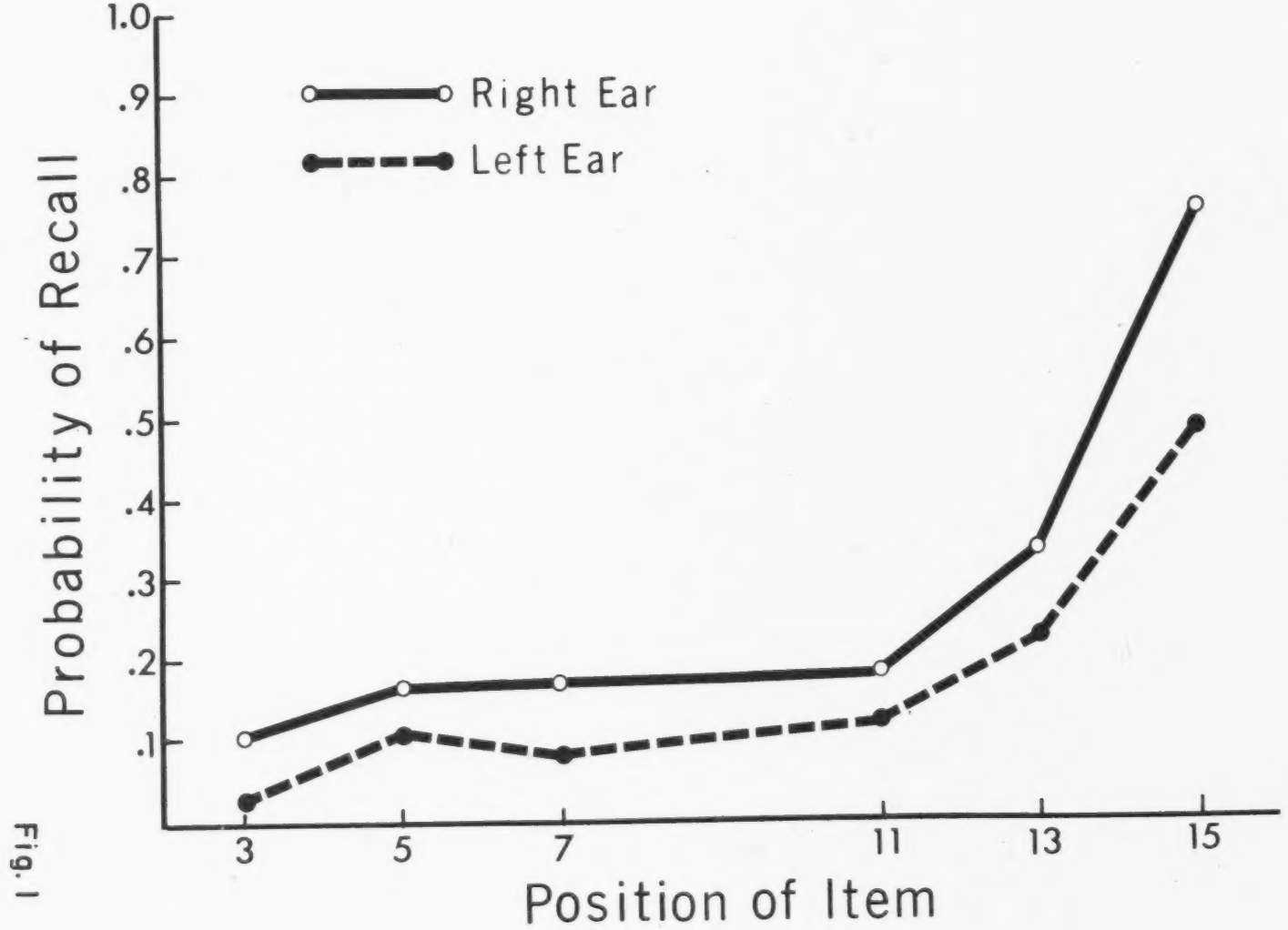


Figure 1: Probability of recalling an item as a function of ear of presentation.

items. Taking the mean recall probability of items in positions 3, 5, and 7 as an estimate of P(LTS), the STS components of the data in Figure 1 can be computed from the equation above uncomplicated by the LTS component. Figure 2 gives the estimated STS and LTS components for right- and left-ear presentation. Wilcoxon matched-pairs tests conducted separately on the STS components and LTS components revealed a right-ear advantage in both cases, $P < .005$ and $P < .001$, respectively.

DISCUSSION

The present experiment has shown that dichotic stimulation is not a necessary condition for demonstrating a right-ear advantage for recall of verbal material. Therefore, we must assume that the crossover auditory pathways are functionally prepotent even in the absence of dichotic stimulation. Thus, an input monaurally to the right ear has some advantage for left-hemisphere processing, and, presumably, an input monaurally to the left ear has an advantage for right-hemisphere processing. When the input is verbal and the left ear is the recipient the major or more important route (but obviously not the only route) taken by that input is the contralateral path to the right hemisphere followed, in turn, by the callosal path to the left hemisphere. On the other hand, a verbal input to the right ear is conveyed more directly to the left hemisphere.

How should we view the advantage of right-ear presentation in this experiment? Let us first examine the ideas forwarded to account for ear asymmetry under conditions of dichotic stimulation. There it has been suggested that the right-ear advantage occurs because the left-ear input traveling an indirect path to the speech processing hemisphere suffers a "loss" of auditory information (Studdert-Kennedy and Shankweiler, 1970). In other words, it is suggested that the left hemisphere receives a comparatively impoverished signal from the left ear for speech processing. The impoverishment may be due to the longer route taken by the left-ear input, i.e., information may be lost during interhemispheric transfer; in addition, or instead, the left-ear input may have to queue because the speech processing machinery to which it seeks access is busily engaged with processing the right-ear input. It is argued that in the course of queueing the left-ear signal decays.

The queueing notion applies to the dichotic situation but not to the poorer left-ear performance in the present experiment. On the other hand, the idea that the left-ear signal is somehow "degraded" during callosal transmission is relevant to the present concern. We could suppose that the degradation which occurs is in the form of a reduced signal-to-noise ratio, or we could argue that what is transmitted callosally is not so much a degraded signal as it is a recoded version of the input, reflecting the processing operation of the right hemisphere. In any event, it can be argued that under conditions of monaural stimulation the speech processing apparatus receives an input from the left ear which is for some reason difficult to process and, therefore, perhaps, takes more time to process than an input from the right ear. This should not be taken to mean that perception of verbal material received by the left ear is less distinct or less adequate than the perception of verbal material received by the right ear. Rather, it is just more difficult for the speech processing machinery to achieve that adequate perception of left-ear speech stimulation.

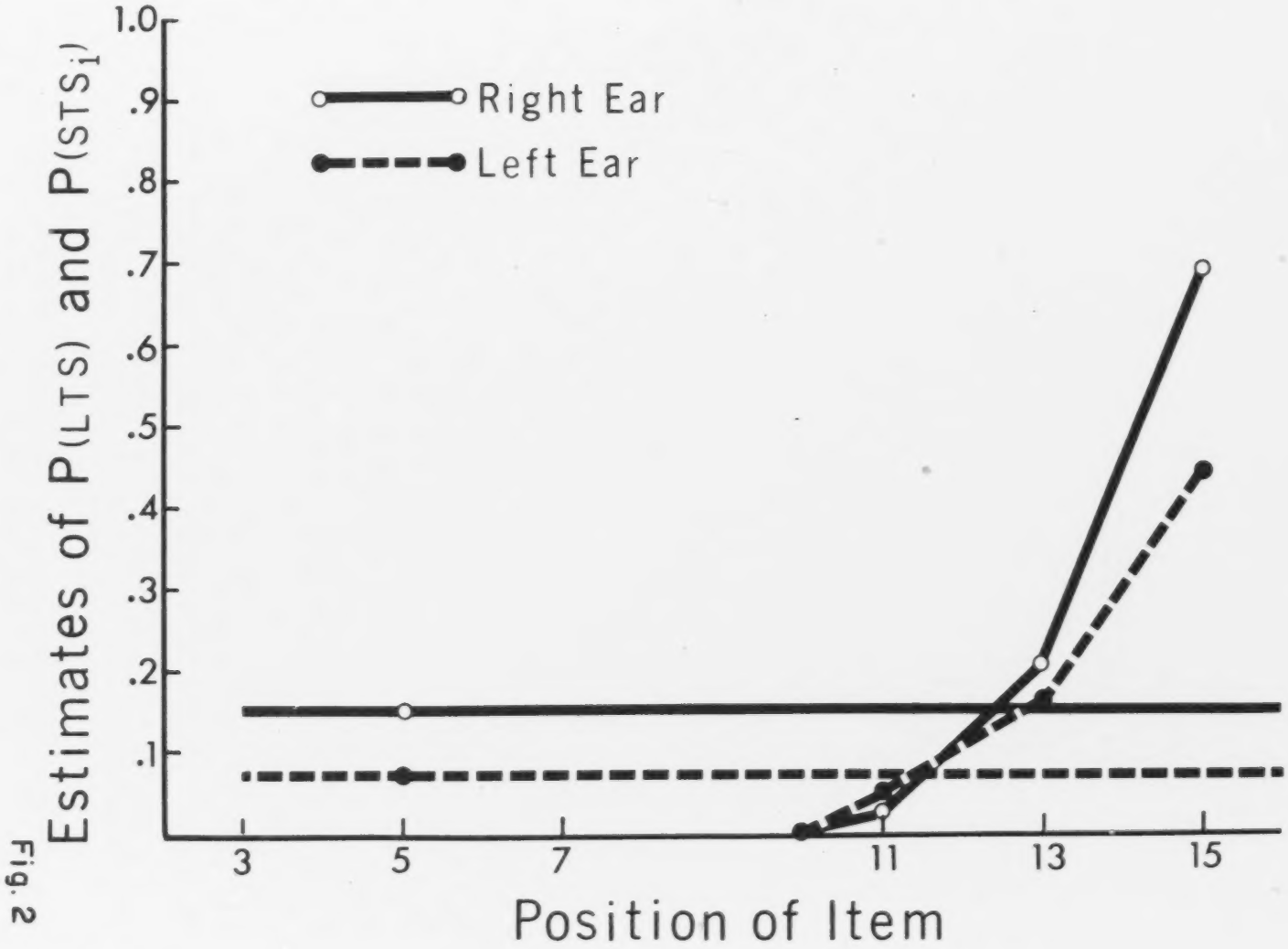


Fig. 2

Figure 2: Estimates of short-term and long-term storage as a function of ear of presentation.

Various experiments have pointed to a close relation between STS and limited processing capacity (e.g., Dillon and Reid, 1969; Posner and Rossman, 1965). If a subject is required to retain verbal items and to perform concurrently a subsidiary arithmetic task, then short-term retention may suffer or performance on the subsidiary task may suffer or both may suffer. In general the more demanding the subsidiary task is, the lower is the level of short-term retention. Both Posner (1966) and Moray (1967) have argued for a limited processing capacity which can be partitioned across the various components of a task or partitioned across the various components of concurrent tasks. If, as argued above, the processing of speech from the left ear is more demanding and more time consuming than the processing of speech from the right ear then presumably less capacity should be available for maintaining left-ear items in STS. Hence, STS for verbal items received on the left ear should be poorer than STS for items received on the right ear. And if items in STS are transferred with some probability to LTS (c.f. Waugh and Norman, 1965) then a poorer STS representation should lead to a poorer LTS representation. Hence, left-ear material should be registered in LTS less accurately than right-ear material.

Central to the foregoing interpretation of the present data is the idea that the processes of perceiving and rehearsing are oppositive. It is suggested that they both compete for the limited central processing capacity. An alternative version of this idea is that the perception and rehearsal of speech share a common device. One major theory holds that speech is perceived by reference to mechanisms underlying articulation (Liberman et al., 1967). It is also a commonplace view that rehearsal of verbal material engages the mechanisms underlying articulation (Peterson, 1969). Thus it can be argued that in a memory task (such as the probe paradigm of the present experiment) both the perception of current verbal items and the rehearsal of earlier ones depend upon the articulatory apparatus. Consequently, where rehearsal demands are high we should expect ear asymmetry in retention to parallel ear asymmetry in perception.

REFERENCES

Bakker, D. J. (1967) Left-right differences in auditory perception of verbal and nonverbal material by children. Quart. J. Exp. Psychol. 19, 334-336.

Bakker, D. J. (1969) Ear-asymmetry with monaural stimulation: Task influences. Cortex 5, 36-42.

Bakker, D. J. and J. A. Boeynga. (1970) Ear order effects on ear-asymmetry with monaural stimulation. Neuropsychologia 8, 385-386.

Bocca, E., C. Caleano, V. Cassinari, and F. Migliavacca. (1955) Testing 'cortical' hearing in temporal lobe tumors. Acta Otolaryng. 45, 289-304.

Broadbent, D. E. and M. Gregory. (1964) Accuracy of recognition for speech presented to the right and left ears. Quart. J. Exp. Psychol. 16, 359-360.

Bryden, M. P. (1963) Ear preference in auditory perception. J. Exp. Psychol. 65, 103-105.

Chaney, R. B. and J. C. Webster. (1965) Information in certain multidimensional signals. U. S. Navy Electronic Laboratories (San Diego, Cal.) Report No. 1339.

Curry, F. K. W. (1967) A comparison of left-handed and right-handed subjects on verbal and nonverbal dichotic listening tasks. Cortex 3, 343-352.

Curry, F. K. W. and P. R. Rutherford. (1967) Recognition and recall of dichotically presented verbal stimuli by right and left handed persons. Neuropsychologia 5, 119-126.

Darwin, C. J. (1969) Auditory perception and cerebral dominance. Unpublished Ph.D. thesis, University of Cambridge.

Dillon, R. F. and L. S. Reid. (1969) Short-term memory as a function of information processing during the retention interval. J. Exp. Psychol. 81, 261-269.

Kimura, D. (1961) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exp. Psychol. 16, 355-358.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-172.

Kintsch, W. (1970) Learning, Memory and Conceptual Processes. (New York: John Wiley).

Liberman, A., F. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Milner, B., L. Taylor, and R. W. Sperry. (1968) Lateralized suppression of dichotically presented digits after commissural section in man. Science 161, 184-185.

Moray, N. (1967) Where is capacity limited? A survey and a model. Acta Psychologica 27, 84-92.

Murphy, E. H. and P. H. Venables. (1970) The investigation of ear asymmetry by simple and disjunctive reaction-time tasks. Percep. Psychophys. 82, 104-106.

Peterson, L. R. (1969) Concurrent verbal activity. Psychol. Rev. 76, 376-386.

Posner, M. I. (1966) Components of skilled performance. Science 152, 1712-1718.

Posner, M. I. and E. Rossman. (1965) Effect of size and location of informational transforms upon short-term retention. J. Exp. Psychol. 70, 496-505.

Rosenzweig, M. R. (1951) Representations of the two ears at the auditory cortex. Amer. J. Physiol. 167, 147-158.

Satz, P. (1968) Laterality effects in dichotic listening. Nature 218, 277-278.

Shankweiler, D. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Sparks, R. and N. Geschwind. (1968) Dichotic listening in man after section of neocortical commissures. Cortex 4, 3-16.

Studdert-Kennedy, M. (in press) The perception of speech. In Current Trends in Linguistics, Vol. XII, ed. by T. A. Sebeok. (The Hague: Mouton).

Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Thorndike, E. L. and I. Lorge. (1944) The Teacher's Wordbook of 30,000 Words. (New York Bureau of Publications: Teachers College, Columbia University).

Waugh, N. C. and D. A. Norman. (1965) Primary Memory. Psychol. Rev. 72, 89-104.

A Right-Ear Advantage in Choice Reaction Time to Monaurally Presented Vowels: A Pilot Study

Michael Studdert-Kennedy[+]
Haskins Laboratories, New Haven

ABSTRACT

In a pilot experiment with five subjects monaural reaction times for the identification of three synthetic, steady-state vowels, /I, ɛ, æ/, were measured as a function of their duration (80 msec vs. 20 msec). On the long (80 msec) vowels, four of the five subjects were faster in identifying vowels presented to their right ears than those presented to their left ears. But the mean right-ear advantage of 22 msec was not significant. On the short (20 msec) vowels all subjects showed a right-ear advantage, and the mean advantage of 48 msec was significant. The shift in the right-ear advantage, as a function of vowel duration, showed small between-subject variability and was highly significant.

That dichotic competition is a necessary condition of the right-ear advantage for the perception of speech has become a commonplace of the literature since Kimura (1961a, 1961b) first formulated the hypothesis and presented evidence in support of it. Her model has subsequently been favored by the results of work with split-brain patients (Milner, Taylor, and Sperry, 1968; Sparks and Geschwind, 1968) and by a number of studies (e.g., Day and Cutting, 1970; Darwin 1971a, 1971b) demonstrating limits on the classes of competing sound sufficient to produce a right-ear advantage. Where such studies show that a competing pure tone, for example, does not produce an ear advantage for certain classes of speech sound, we may reasonably infer that monaural presentation of the same speech sounds would also have failed to yield an ear advantage.

Nonetheless, scattered reports of monaural ear effects have begun to appear. Bakker (1967, 1968, 1969, 1970), for example, has repeatedly shown right-ear advantages for recall of monaurally presented lists of words. Bever (1970); Pisoni, Jarvella, and Tikofsky (1970); and Herman (1972) have reported right-ear advantages in recall, and in time taken for recall, of monaurally presented sentences. Recently, Turvey, Pisoni and Croog (1972) have found right-ear

---

[+]Also Queens College and the Graduate Center of the City University of New York.

Acknowledgment: Thanks are due to Carol Fowler for collecting and analyzing the data.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

advantages in both primary and secondary components of the short-term retention function for monaurally presented word lists. All these studies have dealt with recall rather than with immediate perception. Taken with the assumption that monaural presentation can yield no ear advantage in perception, this leads to a paradox. How can memory for an utterance display an earmark, several seconds after the perceptual process has been completed, if that mark was not imprinted when the information that it conveys was immediately available, namely, during initial perceptual analysis?

The paradox is resolved if we assume that the earmark is imprinted during perception, but that our methods of detecting it have been inadequate. The present pilot study was therefore designed to determine whether a new measure of performance, reaction time, might not reveal monaural ear differences where percentage correct, or its derivatives, had failed. Springer (1971a, 1971b) has demonstrated significant right-ear advantages in reaction time to stop consonants under conditions of dichotic competition, most recently where the competing stimulus was white noise (Springer, 1972), which typically yields no ear effect in measures of percentage correct, and which yielded none in her experiment. She has suggested that reaction time may prove to be a more sensitive measure than gross identification scores in studying laterality effects. For monaural work the use of reaction time seems additionally apt, since the only studies in which monaural ear advantages have been reported (the recall studies mentioned above) are those that have exerted temporal pressure on the perceptual and memorial systems.

As stimuli for the study, steady-state vowels were chosen, partly because they belong to a class of stimuli for which ear advantages in terms of percentage correct have proved difficult to demonstrate, and partly because there is a tempting parallelism between categorical perception studies and dichotic studies. In the first, consonants tend to be perceived categorically, vowels continuously, while in the second, consonants typically yield a right-ear advantage, vowels little or none. Since Pisoni (1971), among others, has demonstrated that perception of brief (50 msec) vowels tends to be more categorical than perception of their longer (300 msec) counterparts, a similar shift in the ear advantage for vowels, as a function of duration, would be a step toward linking the phenomena of categorical perception and right-ear advantages.

METHOD

Two sets of three steady-state vowels, /I, ɛ, æ/, were synthesized on the Haskins Laboratories' parallel resonant synthesizer. The first set had durations of 80 msec (long vowels), the second had durations of 20 msec (short vowels). The intensities of the second set were increased by 6 db relative to those of the first, in order to match the two sets for energy.

For both sets the same test order was used. It consisted of 20 binaural trials in which the three vowels were presented in random order approximately seven times each, and 60 monaural trials in which the three vowels were presented randomly ten times to each ear. For the monaural section, ear of presentation was semi-randomized within blocks of ten so that, in any block, five vowels were presented to the left ear, five to the right. There were 3-sec intervals between trials and a 10-sec interval after every tenth trial.

Five subjects, four women and one man, listened to the tapes individually in a quiet room. They heard each tape twice, once with the earphones reversed to distribute channel effects equally over the two ears. This yielded a total of 40 binaural and 60 monaural trials on each ear for each subject. A subject sat before a reaction time board, on which there were four buttons: a central "home" button, on which the subject rested his right index finger between trials, and three appropriately labeled response buttons arranged above the home button in a 3-inch arc with a radius of two inches. Subjects were instructed to respond as rapidly as they could without losing accuracy.

The stimuli and the subjects' responses were recorded on separate channels of a two-channel tape recorder. Depression of a response button released a characteristic voltage, so that the experimenter could read a VU meter and record the subject's identification. The recorded stimuli and response button voltages were later used to start and stop a Hewlett-Packard electronic counter, yielding reaction times in msec.

RESULTS

Since this was a pilot study, reaction times were not transformed, and means rather than medians were computed. Table 1 presents the mean reaction times in msec for the five subjects under the several conditions of the experiment. The reaction times are relatively long for a three-choice response, but movement from home to response button must account for a fair proportion of the total time.

TABLE 1: Mean reaction times in milliseconds to long (80 msec) and short (20 msec) synthetic vowels presented binaurally and monaurally.

	Long Vowels			Short Vowels		
		Monaural			Monaural	
Subject	Binaural	Left	Right	Binaural	Left	Right
1	641.9	651.3	660.7	753.1	789.0	771.6
2	623.1	688.1	683.7	592.1	693.2	660.8
3	466.7	531.3	519.7	469.5	558.2	512.0
4	675.3	685.5	674.0	632.6	738.9	706.5
5	652.6	782.6	691.8	755.4	822.5	710.8
Mean	611.9	667.8	646.0	640.5	720.4	672.3

As was expected, mean reaction time was longer for the short, phonetically "difficult," vowels than for the long vowels. All subjects show this effect for both ears under monaural presentation. Subjects 2 and 4 reverse the effect under binaural presentation. Binaural reaction times are also faster than monaural, the only exceptions being for subject 4 who is slightly faster on his right ear for the long vowels and for subject 5 who is faster on his right ear for the short vowels.

Table 2 displays ear differences in mean reaction times (RT) to long and short vowels presented monaurally: mean RT for the right ear is subtracted from mean RT for the left ear, so that a positive difference indicates a right-ear advantage, a negative difference a left-ear advantage. For the long vowels

TABLE 2: Ear differences in mean reaction times (RT) in msec to long and short vowels presented monaurally. (L = mean RT for left ear; R = mean RT for right ear).

Subject	Long Vowels L-R	Short Vowels L-R	Short-Long Difference $(L\text{-}R_{short}) - (L\text{-}R_{long})$
1	-9.4	17.4	26.8
2	4.4	32.4	28.0
3	11.6	46.2	34.6
4	11.5	32.4	20.9
5	90.8	111.7	20.9
Mean	21.8	48.0	26.2
t	1.23	2.90	10.28
p	$>.05$	$<.05$	$<.0001$

the differences are fairly small (except for that of subject 5), and one subject (1) shows a left-ear advantage. The mean advantage to the right ear of approximately 22 msec is not significant by a matched pairs t-test. For the short vowels the differences are larger and every subject shows a right-ear advantage: the mean advantage of 48 msec is significant by a matched pairs t-test ($p<.05$).

The between-subject variability in mean ear differences is quite high, largely due to the extreme scores of subject 5. But if we consider the difference between the differences (Table 2, column 3), the variability is strikingly reduced. Every subject gives an increase in right-ear advantage, as a function of vowel duration, of between 20 and 35 msec. The mean increase is approximately 26 msec and a matched pairs t-test, although not independent of the first two tests, yields a t-value sufficiently high for one to be quite confident of its significance ($p<.0001$).

Finally, Table 3 displays the error rates. Reaction times for errors and correct responses were not separated, so that some of the effects reported above could be due to the well-known fact that it takes longer to make a mistake than not to. And indeed the rank order of the mean error rates is almost perfectly correlated with the rank order of the mean reaction times. However, examination of the individual scores suggests that this correlation is not causal, but merely a reflection of the fact that the right ear is superior in both accuracy and speed. Subject 1, for example, whose performance is virtually error-free shows essentially the same pattern of RT differences as Subject 2, whose error rates shift from a left-

TABLE 3: Percentages of errors for long (80 msec) and short (20 msec) synthetic vowels presented binaurally and monaurally.

	Long Vowels			Short Vowels		
		Monaural			Monaural	
Subject	Binaural	Left	Right	Binaural	Left	Right
1	0	2	2	0	0	0
2	0	7	10	10	12	7
3	5	7	2	5	32	22
4	2	0	0	7	22	17
5	7	15	7	0	7	5
Mean	3	6	4	4	14	10

ear advantage on the long vowels, to a right-ear advantage on the short. Subject 5, whose monaural error rate on the long vowels was higher for both ears than on the short, not only gives faster reaction times on the long vowels, but also displays exactly the same shift in mean RT difference, as a function of vowel duration, as subject 4 whose monaural error rates move from zero on the long vowels to around 20 percent on the short. Thus, while error responses will have to be segregated from correct responses in a full experiment, their integration in the present study does not seem to underlie the observed RT differences.

DISCUSSION

If the results of this pilot study are borne out by later work, they will have wide implications for the study of speech laterality effects. Substantively, they represent the first demonstration of a right-ear advantage for vowels, as a function of their stimulus properties rather than their experimental context (cf. Haggard, 1971; Darwin, 1971b). A previous experiment with short (40 msec) vowels, using a percentage correct measure, showed no right-ear advantage (Darwin, 1969), and the lack of a reliable ear advantage for vowels has been the premise of a good deal of theorizing (e.g., Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Haggard, 1971).

Theoretically, the results provide the glimmer of a rational basis for understanding monaural ear effects in recall. If the right-ear/left-hemisphere system has a temporal advantage in perception of a single phonetic segment, it is not surprising that, as time pressure increases over a relatively long sequence of segments and reduces the opportunity for rehearsal, an initial temporal advantage should ultimately emerge as an advantage in terms of percentage correct. This view may, furthermore, serve to link the phenomena of categorical perception and laterality, hinting at their common origin in the engagement of a phonetic processing mechanism, specialized for rapid response.

Methodologically, the results may simplify the experimental analysis of the ear advantage. There is no question that dichotic competition serves to magnify observed ear advantages and that dichotic experiments may, by generating a sizable number of errors for analysis, permit theoretical decomposition of the per-

ceptual process. At the same time, monaural reaction time procedures could serve to reexamine, in relatively short order, the classes of stimuli for which ear advantages can be demonstrated.

REFERENCES

Bakker, D. J. (1967) Left-right differences in auditory perception of verbal and nonverbal material by children. Quart. J. Exp. Psychol. 19, 334-336.

Bakker, D. J. (1968) Ear asymmetry with monaural stimulation. Psychon. Sci. 12, 62.

Bakker, D. J. (1969) Ear asymmetry with monaural stimulation: Task influences. Cortex 5, 36-41.

Bakker, D. J. (1970) Ear asymmetry with monaural stimulation: Relations to lateral dominance and lateral awareness. Neurophychologia 8, 103-117.

Bever, T. G. (1970) The nature of cerebral dominance in speech behavior of the child and adult. In Mechanisms of Language Development, ed. by Huxley and Ingram. (New York: Academic Press).

Darwin, C. J. (1969) Auditory perception and cerebral dominance. Unpublished Ph.D. thesis, University of Cambridge.

Darwin, C. J. (1971a) Dichotic forward and backward masking of speech and non-speech sounds. J. Acoust. Soc. Amer. 50, 129(A).

Darwin, C. J. (1971b) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 46-62.

Day, R. S. and J. E. Cutting. (1970) Perceptual competition between speech and nonspeech. J. Acoust. Soc. Amer. 49, 85(A). (Also in Haskins Laboratories Status Report on Speech Research SR-24, 35-46.)

Haggard, M. P. (1971) Encoding and the REA for speech signals. Quart. J. Exp. Psychol. 23, 34-45.

Herman, S. (1972) The right ear advantage for speech processing. Paper presented at 84th meeting of the Acoustical Society of America, Miami Beach, Fla., November.

Kimura, D. (1961a) Some effects of temporal-lobe damage on auditory perception. Canad. J. Psychol. 15, 156-165.

Kimura, D. (1961b) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Milner, B., L. Taylor, and R. W. Sperry. (1968) Lateralized suppression of dichotically presented digits after commissural section in man. Science 161, 184-185.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. thesis, University of Michigan. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Pisoni, D. B., R. J. Jarvella, and R. S. Tikofsky. (1970) Laterality factors in the perception of sentences varying in semantic constraint. J. Acoust. Soc. Amer. 47, 76(A).

Sparks, R. and N. Geschwind. (1968) Dichotic listening in man after section of neocortical commissures. Cortex 4, 3-16.

Springer, S. P. (1971a) Ear asymmetry in a dichotic listening task. Percept. and Psychophys. 10, 239-241.

Springer, S. P. (1971b) Lateralization of phonological processing in a dichotic detection task. Unpublished Ph.D. thesis, Stanford University.

Springer, S. P. (1972) The effect of contralateral noise on the perception of speech: A reaction time analysis. Paper presented to the meeting of the Western Psychological Association, Portland, Ore., April.

Turvey, M. T., D. B. Pisoni, and J. F. Croog. (1972) A right-ear advantage in the retention of words presented monaurally. Haskins Laboratories Status Report on Speech Research SR-31/32 (this issue).

Perceptual Processing Time for Consonants and Vowels*

David B. Pisoni+
Haskins Laboratories, New Haven

ABSTRACT

Perceptual processing time for brief CV syllables and steady-state vowels was examined in a backward recognition masking paradigm. Subjects were required to identify a 40 msec sound selected from either a consonant set (/ba/, /da/, /ga/) or a vowel set (/i/, /I/, /ɛ/). The target sound was followed by a different sound drawn from the same set after a variable silent interstimulus interval. The second sound interrupted the perceptual processing of the target sound at short interstimulus intervals. Recognition performance improved with increases in the silent interstimulus interval. One experiment examined processing time for consonants and vowels under binaural presentation. Two additional experiments compared consonant and vowel recognition under both binaural and dichotic presentation. The results indicated that: (1) consonants require more processing time for recognition than vowels and (2) binaural and dichotic presentation conditions produce differential effects on consonant and vowel recognition. These findings have several important implications for understanding the recognition process. First, speech perception is not immediate, but is the result of several distinct operations which are distributed over time. Second, speech perception involves various memorial processes and mechanisms which recode and store information at different stages of perceptual analysis.

One of the most basic questions in speech perception concerns the process of recognition. How is a particular speech sound identified as corresponding to a specific phonetic segment? Although many of the current theories of speech perception have focused on the recognition process for some time, they have all been quite vague in their approach to this problem (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Stevens and House, 1972). It is

---

*Paper presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., 1 December 1972. The research was supported by a grant from NICHD to Haskins Laboratories and a PHS grant (SO5 RR 7031) to Indiana University.

+Also Indiana University, Bloomington.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

usually assumed that the recognition process entails a series of stages and operations in which the acoustic stimulus makes contact with either a stored representation in long term memory or some representation that may be constructed or generated from rules residing in long term memory (Liberman et al., 1967; Halle and Stevens, 1962). Unfortunately, little empirical work has been directed at examining these hypothetical stages or specifying what types of operations might be involved in the recognition process for speech sounds.

The present study is concerned with mapping out in a quantitative way the earliest stages of perceptual processing for speech sounds. To achieve precise control over early processing, a backward recognition masking procedure was used. With this technique the processing of one stimulus may be interrupted at various times after its presentation by another stimulus and thereby provide information about the temporal course of perceptual processing (Massaro, 1971, 1972).

Figure 1 shows the general features of the backward recognition masking paradigm used in the present series of experiments. On each trial the listener is presented with two successive stimuli but is required to identify only the first stimulus or target sound. The second sound in the sequence serves as the masking stimulus and is presented after some variable silent interstimulus interval. When the mask follows the target at very short intervals it may interrupt or interfere with the processing of the target sound. By varying the duration of the silent interval between the target and mask it is possible to determine the amount of processing time needed for recognition of the target sound. The perceptual processing time for the recognition of brief consonant-vowel (CV) syllables and steady-state vowels was examined in this study because consonants and vowels not only differ in their acoustic properties but have also been shown to have basically different perceptual characteristics (Liberman et al., 1967; Studdert-Kennedy and Shankweiler, 1970).

METHOD

The stimulus conditions employed in these experiments are shown in Figure 2. The consonant stimuli were the CV syllables /ba/, /da/, and /ga/. They were 40 msec in duration and had formant transitions lasting 20 msec. The steady-state vowels used were /i/, /I/, and /ɛ/ and they were also 40 msec in duration. The target sound was selected from either the consonant set or the vowel set. A given target stimulus was then followed by a different stimulus drawn from the same set after a variable silent interstimulus interval. The three sounds within each stimulus condition were arranged in all possible permutations to produce the six stimulus pairs shown in Figure 2. Each pair represented a target and masking sound combination. The intensity relations between target and mask were also manipulated but they will not be discussed here since the effects are not relevant to the major conclusions.

The details of the experimental design are shown in Figure 3. In each of the experiments the intensity and interstimulus interval variables were identical and completely random across trials. Experiment 1 compared consonant and vowel recognition under binaural presentation conditions with the same group of listeners. Experiments 2 and 3 compared binaural and dichotic masking conditions for consonants and vowels with separate groups of listeners. In the

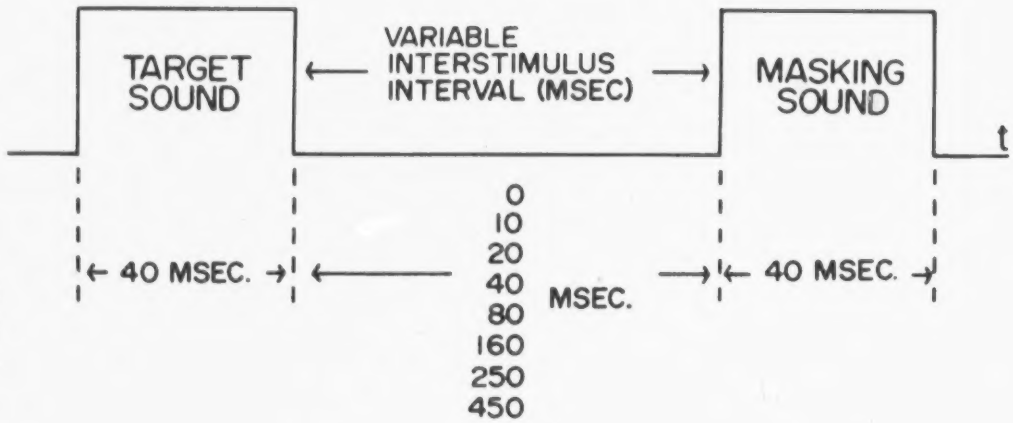


Fig. 1

Figure 1: Description of the backward recognition masking paradigm used in the present experiments.

STIMULUS CONDITIONS

I. CONSONANTS VS VOWELS

CONSONANT - VOWEL SYLLABLES (40 MSEC)	STEADY - STATE VOWELS (40 MSEC)
BA - DA - GA	i - I - ɛ
STIMULUS PAIRS:	STIMULUS PAIRS:
1) BA - DA — DA - BA	1) i - I — I - i
2) BA - GA — GA - BA	2) i - ɛ — ɛ - i
3) DA - GA — GA - DA	3) I - ɛ — ɛ - I

II. INTENSITY RELATIONS: $\Delta I = I_{TARGET} - I_{MASK}$

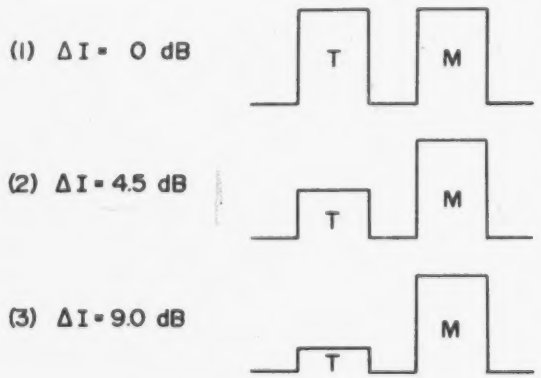


Fig. 2

Figure 2: Description of the stimulus conditions used in the present experiments.

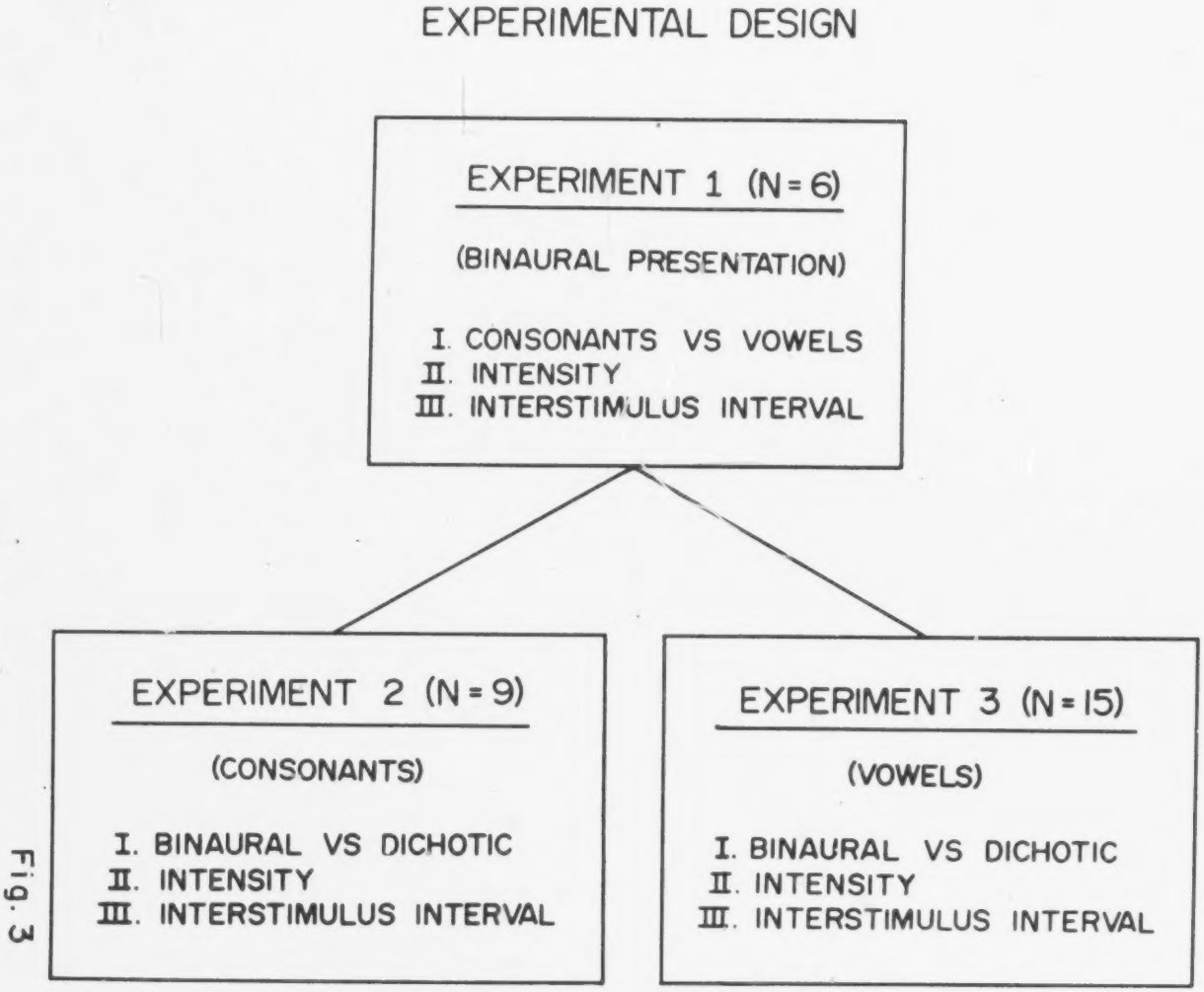


Fig. 3

Figure 3: Experimental design of the three experiments.

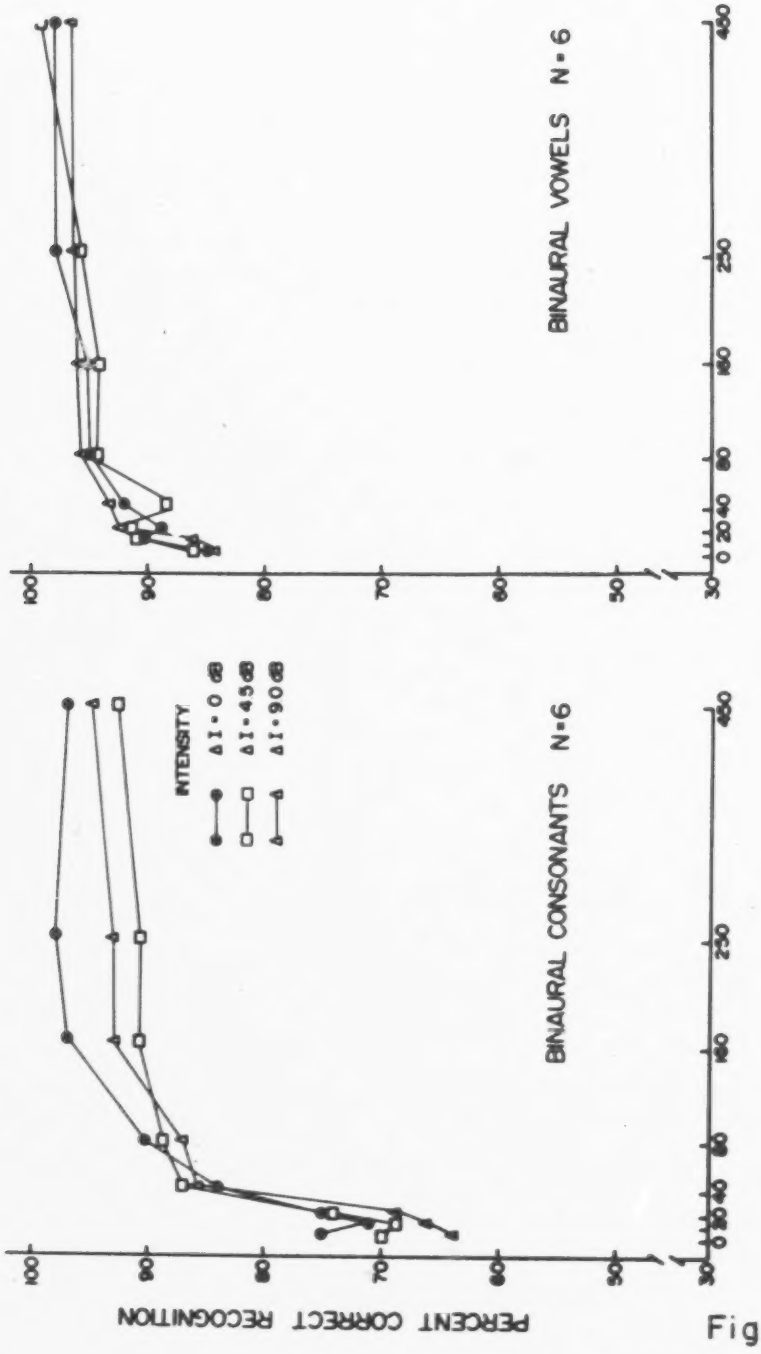


Fig. 4

Figure 4: Average recognition scores for consonants and vowels in Experiment 1 as a function of interstimulus interval.

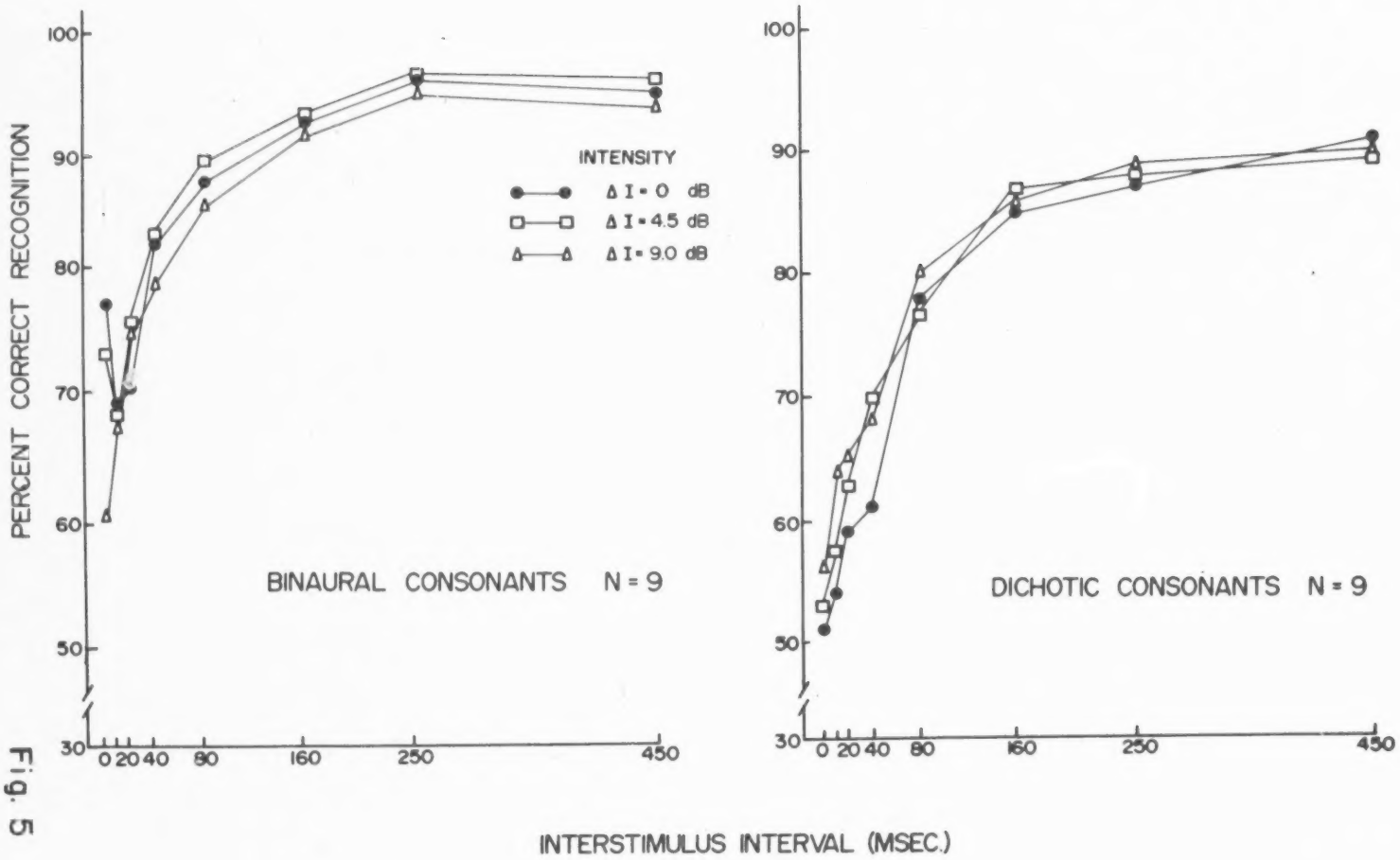


Fig. 5

Figure 5: Average recognition scores for consonants under binaural and dichotic presentation conditions in Experiment 2 as a function of interstimulus interval.

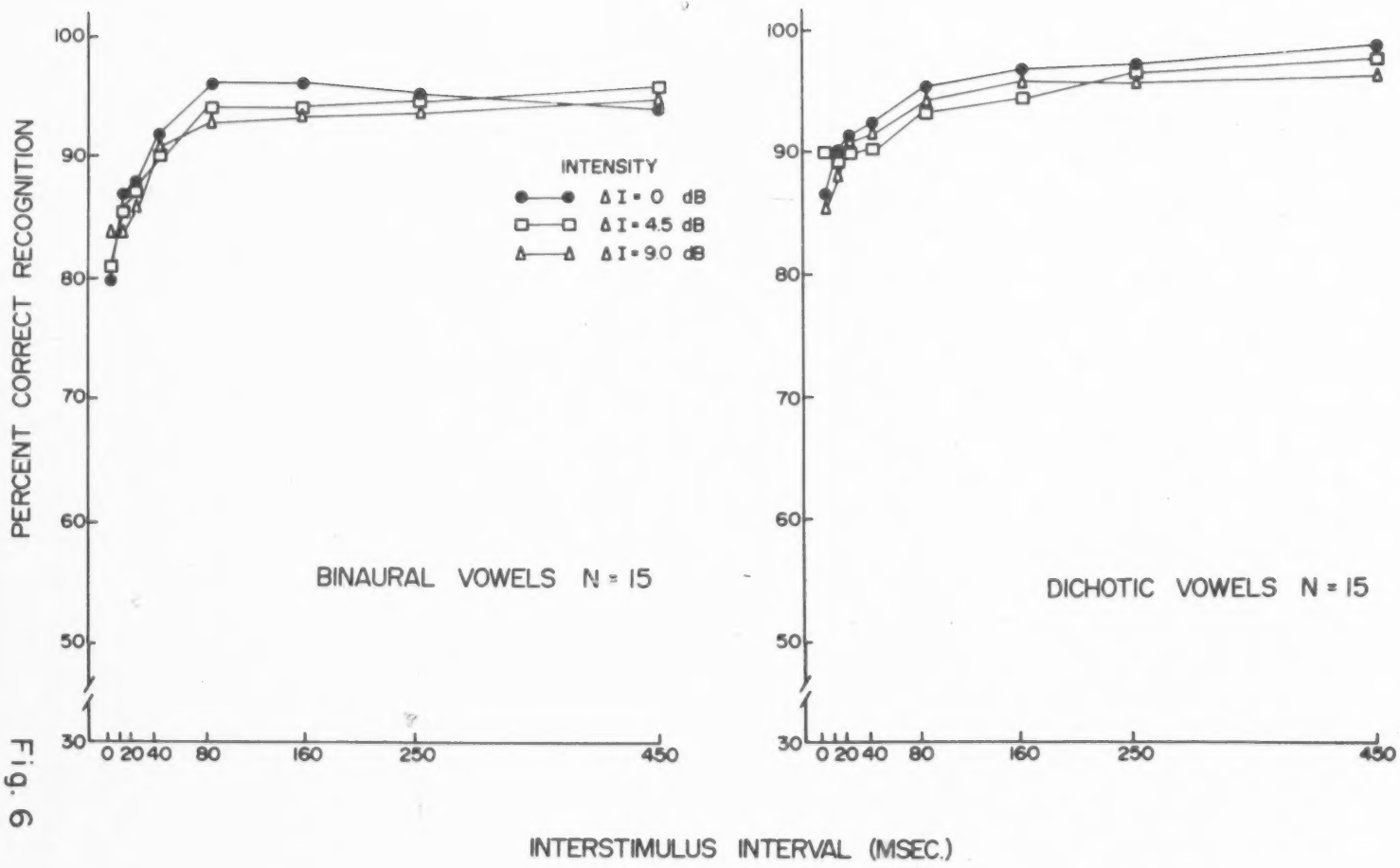


Fig. 6

Figure 6: Average recognition scores for vowels under binaural and dichotic presentation conditions in Experiment 3 as a function of interstimulus interval.

binaural conditions target and mask were presented to both ears. In the dichotic condition the target sound was presented to one ear and the mask was presented to the other. Targets and masks were presented equally often to both ears in the dichotic condition.

RESULTS AND DISCUSSION

Figure 4 shows the results of Experiment 1 which compared consonant and vowel stimuli under binaural presentation. Recognition performance is expressed in terms of the percent correct identification of the target sound. Recognition improves with increases in the silent interstimulus interval for both vowels and consonants. However, the masking appears to be much more effective for the consonants than the vowels. This is especially noticeable at short interstimulus intervals and indicates that consonants need more time for recognition than vowels.

The results of the second experiment which examined processing time for the consonant stimuli under binaural and dichotic presentation conditions are shown in Figure 5. The binaural data are almost identical to the findings obtained in the first experiment. However, there is a large and consistent difference between the dichotic and binaural presentation conditions. Performance under dichotic presentation is lower overall than performance under binaural presentation. Moreover, performance under dichotic presentation at short intervals appears to be markedly different than performance under binaural presentation.

The results of the third experiment which studied recognition of vowels under binaural and dichotic presentation conditions are shown in Figure 6. The effect observed in Figure 5 for the consonants is strikingly absent. There is again some masking at short intervals for the vowels but there is no difference between binaural and dichotic presentation conditions.

In summary, when the target stimulus was followed by the masking sound at short interstimulus intervals recognition of the target was interrupted, suggesting that perceptual processing for speech sounds continues even after the stimulus has ended. This finding indicates that speech perception is not a result of immediate stimulation but rather requires a certain amount of processing time for the extraction of relevant features from the acoustic signal.

The present results also reveal that consonants require more processing time for recognition than vowels. Additionally, when binaural and dichotic masking conditions were compared for these classes of speech sounds differences in recognition were obtained for consonants but not for vowels. This last result suggests that there may be an additional stage or stages of perceptual processing needed for consonant recognition that is not needed for vowel recognition.

REFERENCES

Halle, M. and K. N. Stevens. (1962) Speech recognition: A model and a program for research. IRE Transactions on Information Theory IT-8, 155-159.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Massaro, D. W. (1970) Preperceptual auditory images. J. Exp. Psychol. 85, 411-417.

Massaro, D. W. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.

Stevens, K. N. and A. S. House. (1972) Speech perception. In Foundations of Modern Auditory Theory, ed. by J. Tobias. (New York: Academic Press).

Studdert-Kennedy, M. and D. P. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

A Preliminary Report on Six Fusions in Auditory Research

James E. Cutting[+]
Haskins Laboratories, New Haven

Since the 1950's a number of auditory phenonema have been called "fusion" by various researchers. Broadbent (1955), Day (1968), and Halwes (1969), among others, have described experimental situations in which two auditory signals are perceived as one. From the titles of their papers one would assume that they are concerned with the same process: "On the fusion of sounds reaching different sense organs" (Broadbent and Ladefoged, 1957); "Fusion in dichotic listening" (Day, 1968); and "Effects of dichotic fusion on the perception of speech" (Halwes, 1969).

However, fusion is not just one phenomenon, but many phenomena which are, at best, only tenuously related. Subsuming them all under the single label "fusion" with no descriptive adjective easily leads to confusion. The purpose of this paper is to act as a preliminary report, delimiting the various types of auditory fusion, investigating their similarities and dissimilarities, and arranging them on a cognitive hierarchy according to the processing characteristics in each. We will consider six different types of fusion, beginning with the more primitive fusion phenomena and moving towards more complex phenomena.

Cognitive Levels and Criteria for Fusion Classifications

Before considering the various types of fusion, it is necessary to define the levels which we will discuss and to establish criteria for judging the placement of fusions at these levels. We will consider two general levels: designated "higher" and "lower" levels.

Lower-level fusions are energy-dependent. Pitch and intensity are examples of energy parameters. When lower-level fusions are involved small differences in pitch (2 Hz) or small differences in intensity (2 db) between the two to-be-fused stimuli may inhibit fusion or change the fused percept. Timing, in terms of the relative onset time of the two stimuli, is another important parameter. If one stimulus precedes the other by a sufficient interval, fusion no longer occurs and two stimuli are heard. This interval is very small for lower-level fusions, often a matter of microseconds (microsec).

Higher-level fusions are energy-independent. Pitch and intensity may vary between the two stimuli within a much greater range in the higher-level fusions. Differences in the stimuli of 20 Hz or 20 db may not inhibit fusion at all.

[+]Also Yale University, New Haven.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

Relative onset time also plays a lesser role; the two stimuli may vary in relative onsets within a range well beyond that of the lower-level fusions. Higher-level fusions are often insensitive to differences of from 25 to 150 msec. Information, not energy or timing, seems to be important in these fusions, and, as we shall see later, information in the to-be-fused stimuli can often override energy differences.

Other, more psychological characteristics also distinguish the higher fusions from lower fusions. Lower-level fusions are generally passive phenomena, whereas higher-level fusions are more constructive in nature. In the lower-level fusions the subject listens to one clear auditory image and is usually unaware that two stimuli are being presented. In the higher-level fusions, however, the subject listens to a more diffuse auditory image; he may even report hearing two sounds or one sound that sounds a bit strange.

In all six types of fusion, the to-be-fused stimuli are presented dichotically, one stimulus to the right ear and the other stimulus to the left ear. In each case the subject is asked to report what he heard. A diagram of each type is shown in schematic form in Figure 1; they will be discussed in turn below. There is a temptation to think of all six fusions primarily as central processes; stimuli transmitted by different channels and integrated into a single percept. This judgment may be misleading. In vision research, Turvey (in press) has noted that peripheral processes tend to be those which are affected by changes in the energy of the stimuli, while central processes tend to be those which are independent of stimulus energy and are more concerned with the information in the stimuli. This peripheral/central dichotomy parallels the lower-level/higher-level distinction outlined above for auditory fusions. If lower-level fusions are energy-dependent, perhaps they are primarily concerned with peripheral mechanisms. If, on the other hand, higher-level fusions are energy-independent, perhaps they are primarily concerned with central mechanisms. For the purposes of this paper peripheral and central processes will be synonymous with lower and higher cognitive levels.

1. SOUND LOCALIZATION: Fusion of two identical events.

Sound localization has been included as a form of fusion to give a reference point in considering other types of fusion. All audible sounds, simple or complex, can be localized--and usually are. It is the most basic form of auditory fusion and occurs for both speech and nonspeech sounds. The best way to study sound localization in the laboratory is to use the same apparatus needed for studying other types of fusion: a good set of earphones, a dual-track tape recorder, and a two-channel tape with appropriate stimuli recorded on it.

Three parameters affect sound localization: pitch, intensity, and timing. First, consider pitch. If two tones are presented, one to each ear, the subject may fuse (localize) them. If the tones have the same pitch, fusion occurs and one tone is heard. If the tones differ by 2 Hz, fusion begins to disintegrate and a more wavering tone is heard. Differences of more than 2 Hz often inhibit fusion altogether, and two tones are heard.[1]

---

[1]This range is particularly relevant for tones below 1000 Hz. Above 1000 Hz the effect is produced by slightly larger pitch differences.

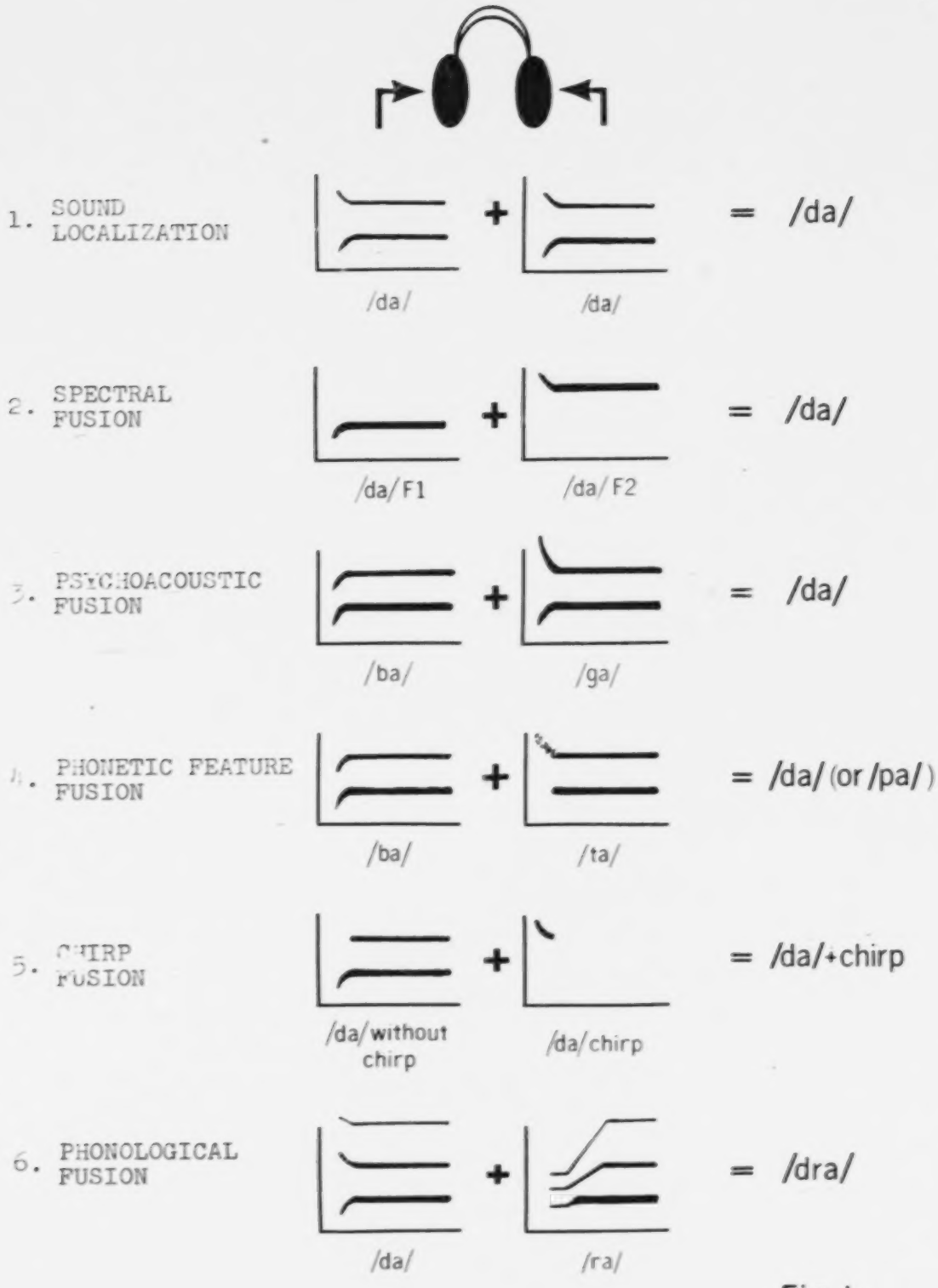


Figure 1: Six fusions of /da/. Schematic spectrograms of speech and speech-like stimuli in six types of auditory fusion.

Intensity is a second parameter which affects localization. If we present two pure tones, one to each ear, at the same frequency and the same intensity the subject usually reports that the apparent source of the tone is "in the middle of the head," or at the midline. If one tone is increased by 2 db without changing the intensity of the other stimulus, he normally reports that the sound has moved in the direction of the ear that received the more intense stimulus. The more we increase the intensity of the louder stimulus, the more the apparent source moves away from the midline and towards the ear with the more intense sound.

The third parameter which affects sound localization is timing. If we present a brief click simultaneously to each ear, the subject reports hearing one click localized at the midline. Delaying one click by 500 microsec causes the apparent source of the click to move away from the midline toward the ear with the leading click. If we delay one click by only 1 msec, the apparent source moves far enough from the midline so that the subjective percept is that one click was presented to one ear with nothing presented to the other. With delays of longer than about 2 msec fusion disintegrates and two clicks are heard. Another form of timing differences also affects localization. If two tones are presented, one to each ear, at the same pitch and intensity, but one tone is slightly delayed, the two tones will be out of phase. In other words, the two ears receive different aspects of the same waveform at a given point in time. If the delay is about 500 microsec the apparent source of the sound moves away from the midline towards the ear that received the leading sound. This perception is produced by phase differences in the stimuli, not by relative onset differences. The fact that localization is highly sensitive to energy parameters of the stimuli and is intolerant of small differences in stimulus timing suggests that it is a lower-level process.

Both speech and nonspeech sounds are fused in sound localization. The first display in Figure 1 shows an example of the localization of speech sounds. If /da/ is presented to both ears at the same time, at the same intensity, and with the same fundamental pitch, a single /da/ will be perceived by the subject at the midline. The same result occurs when a nonspeech sound replaces the /da/ stimuli.

2. SPECTRAL FUSION: Fusion of different spectral parts of the same signal.

Broadbent (1955) and Broadbent and Ladefoged (1957) reported a second type of fusion. Spectral fusion occurs when different spectral ranges of the same signal are presented to opposite ears. A given stimulus is filtered into two parts: one containing the low frequencies and one containing the high frequencies. Each is then presented separately but simultaneously to a single ear. The subject reports hearing the original stimulus, as if it had undergone no special treatment. In his initial study Broadbent found that fusion readily occurred for many complex sounds. Subjects fused metronome ticks and fused certain speech sounds when these stimuli were filtered and presented to opposite ears. When the subjects were informed about the nature of the stimuli and asked to report which ear had the low frequency sounds and which ear had the high frequencies they performed at chance level.

Pitch is an important parameter in spectral fusion. Just as there is no sound localization for dichotic tones of different pitches, there is no spectral fusion of complex dichotic stimuli with different pitches (fundamental frequencies). Broadbent and Ladefoged (1957) found that the fundamental frequencies of

the to-be-fused stimuli must be identical for fusion to occur. They presented the first formant of a steady-state vowel to one ear, and the second formant to the other. Subjects fused the stimuli when they had the same pitch, but did not fuse when they had different pitches. Instead, subjects heard two nonspeech sounds. Halwes (1969) found that fundamental frequency differences of 2 Hz are sufficient to inhibit this type of lower-level fusion.

The fundamental frequency of a speech stimulus is analogous to the carrier-band frequency of a radio station signal. If the to-be-fused spectral stimuli have different pitches, their entire waveforms are as different as if they had been broadcast by two different radio stations; the sound waves of the stimuli are not of the same duration, or even multiples of one another. Like a radio set which cannot integrate (receive) two different radio stations at once, the subject cannot integrate (fuse) two spectral stimuli with different pitches. The information that the two stimuli are the formants appropriate for the perception of a vowel cannot overcome pitch disparities in the signals, and no fusion occurs.

The effect of intensity differences between the stimuli has not been explored systematically. Nevertheless, preliminary investigations indicate that spectral fusion is quite sensitive to small differences in intensity between the to-be-fused stimuli.

When the relative onset time of the high and low frequency components of the same signal is altered beyond a few msec fusion no longer occurs and the subject hears two separate signals. For example, when the different spectral parts of metronome ticks are offset by as little as 5 msec, the subject hears two sets of ticks, not one. If filtered speech passages begin at slightly different times, the subject reports hearing two speech-like passages, not one. Thus, spectral fusion is very sensitive to small changes in timing.[2]

Speech sounds more complex than steady-state vowels are also subject to spectral fusion. The second display of Figure 1 shows the first formant of /da/ presented to one ear, and the second formant to the other ear. Provided these stimuli have the same pitch the subject will report hearing one stimulus, the syllable /da/.

---

[2]There is, however, an exception to this timing sensitivity. Using the filters explained previously, Broadbent (1955) presented the first formant of the steady-state vowel /i/ as in BEET to one ear and the second and third formants to the other. The vowel sound was continuous over a duration of many seconds, with a constant pitch, and was recorded on a tape loop. One part of the loop was presented to one ear and a different part of the loop to the other, with the relative timing of the filtered segments off by as much as two or three seconds. Yet fusion occurred: the subjects heard /i/. The explanation for this result is quite simple. Timing is not important in a steady-state vowel with a constant pitch; unlike most other speech sounds, one section is exactly like any other section. Therefore, one section of the filtered vowel fuses with any other appropriately filtered section. This may be the exception which proves the rule that spectral fusion is highly sensitive to timing differences in the stimuli.

3. PSYCHOACOUSTIC FUSION: Fusion of one feature by acoustic averaging.

Psychoacoustic fusion is a third type of fusion which occurs at lower levels of cognitive processing. Although it probably occurs for both speech and nonspeech stimuli, the examples considered here involve speech stimuli. Unlike the stimuli in previously discussed fusions, the stimuli in psychoacoustic fusion have different information (in terms of phonetic features) in the same spectral range.

To demonstrate psychoacoustic fusion, let us choose two consonant-vowel (CV) stimuli, /ba/ and /ga/. Both are restricted to two formants, and parameters such as pitch, formant frequencies, and duration must be identical. The only difference between the stimuli are the direction and extent of the second formant (F2) transition. As shown in the third display of Figure 1, if /ba/ is presented to one ear and /ga/ to the other, the most frequent error is /da/ (Halwes, 1969:61). Halwes found that in such a situation the subject often reports hearing one of the given stimuli, /ba/ or /ga/; however, when he does make an error it usually is the fusion /da/.

Given the stimuli /ba/-/ga/, how does /da/ result? The stimuli differ in terms of the F2 transitions. However, the remainder of both stimuli are identical and are localized at the midline. The two F2 transitions, one rising in /ba/ and one falling in /ga/, appear to be algebraically averaged in such a manner that the subject perceives a stimulus with an intermediate F2 transition, /da/. (Note that an analogous situation arises when the phonemes /p/ and /k/ are in competition: /t/ is often perceived.)

Psychoacoustic fusion appears to be a lower-level, peripheral process because, as in the previously discussed phenomena, fusion occurs only when the pitch of the two stimuli is identical. Preliminary studies show that if the pitch of the two stimuli, /ba/ and /ga/, differs by as little as 2 Hz, the subject rarely reports hearing /da/, thus indicating that no acoustic averaging takes place.

The effects of intensity and timing differences between the stimuli have not been systematically studied. Nevertheless, pilot work suggests that psychoacoustic fusion is sensitive to small stimulus differences in both parameters.

We will reconsider psychoacoustic fusion after we have considered phonetic feature fusion. The two processes are similar yet different, in many ways. Appropriate comparisons will be made between the two.

4. PHONETIC FEATURE FUSION: Fusion of two features by phonetic blending.

With this fourth type of fusion we move to a more central, higher-level process. This is not to say that peripheral mechanisms are inactive, but these fusions cannot be explained wholly by such mechanisms. Halwes (1969) and Studdert-Kennedy and Shankweiler (1970) have reported that phonetic "blending" occurs in the dichotic competition of stop-vowel syllables. This "blending" is phonetic feature fusion. In the fourth display of Figure 1, we note that when the syllable /ba/ is presented to one ear and the syllable /ta/ to the other ear, the subject often reports hearing a syllable which was not presented. The most frequent errors are the "blends" /da/ and /pa/. Here, the subject appears to combine the voicing feature of one of the stimuli with the place feature of

the other. For example, the voicing feature of /b/ is combined with the place feature of /t/ and the result is the fusion /d/.

Let us assume a stimulus repertory of six items: /ba, da, ga, pa, ta, ka/. On a particular trial, three types of responses may occur: correct responses, "blend" errors, and anomalous errors. Given a stimulus pair such as /ba/-/ta/, "blend" errors are much more common than anomalous errors. The anomalous errors for this example are /ga/ and /ka/; both share a voicing feature with one of the pair but they do not share the place feature with either. Using natural speech stimuli Studdert-Kennedy and Shankweiler (1970) found that the ratio of "blend" errors to anomalous errors was 2:1, a rate significantly greater than chance.

Phonetic feature fusion is more central than the fusions discussed above because, for the first time in this discussion, the subject fuses even when the stimuli do not have the same pitch. Using synthetic stimuli Halwes (1969) reports that "blend" errors occur almost as frequently when the two competing stimuli have different pitches as when they have the same pitch.

The effect of intensity and timing differences between the stimuli have not been systematically studied in phonetic feature fusion. However, we may draw on experimental evidence from other sources. Thompson, Stafford, Cullen, Hughes, Lowe-Bell, and Berlin (1972) have noted that dichotic competition occurs for stop-vowel syllables even when one stimulus is 30 db louder than the other. When stimuli such as /ba/ and /ta/ compete phonetic feature fusion may occur, even with such large intensity differences. We will reconsider the nature of dichotic competition and perceptual fusion later in this discussion.

Evidence concerning the effect of timing differences in phonetic feature fusion is also indirect. Studdert-Kennedy, Shankweiler, and Schulman (1970) have shown that when two CV syllables are presented at various relative onsets, the subject tends to report the syllable which began second better than the syllable which began first. This tendency is very pronounced between relative onset differences of 25 to 70 msec, and has been called the "lag effect." In this region the first stimulus appears to be masked by the second stimulus. If such masking occurs, fusion cannot occur because the phonetic information in the first stimulus is lost. Thus, from these results, we may assume that phonetic feature fusion occurs for temporal onset differences of up to about 25 msec, but that beyond 25 msec the "lag effect" inhibits any fusion that might occur.

Recent evidence indicates that parameters other than pitch, intensity, and timing may differ between the stimuli, and fusion rate still remains the same. For example, the vowels of the stimuli may be different, and fusion still occurs. The data of Studdert-Kennedy, Shankweiler, and Pisoni (1972) show that phonetic feature fusion occurs almost as readily when the stimuli are /bi/-/tu/ as when they are /bi/-/ti/.[3] In both cases the subject is likely to respond with syllables beginning with the "blend" phonemes /d/ or /p/.

A comparison of psychoacoustic fusion and phonetic feature fusion. Because psychoacoustic fusion (fusion type 3) and phonetic feature fusion (fusion type 4) are highly confusable, it is important to make direct comparisons between them. Both processes involve the simultaneous presentation of phonetic

---

[3] D. B. Pisoni, personal communication.

information, and both result in general information loss. In the case of psychoacoustic fusion the /b/ and /g/ features merge into a /d/ feature, and the /b/ and /g/ are lost. In the case of phonetic feature fusion the subject hears two stimuli, and when he does not perceive both of them correctly, he tends to "blend" them. In the process of listening to the dichotic pair, the features associated with a particular stimulus appear to lose their source. The /b/ in /ba/ ceases to be a /b/ and becomes a series of features which are appropriate to /b/: stop consonant, voiced, labial. When another stop consonant competes with /b/, the organization of these features appears to be disrupted, and they frequently get inappropriately reassigned to those of the competing stimulus. Thus, /ba/ plus /ta/ may become /da/, and the /b/ and /t/ are lost.

Dimensions which distinguish the two types of fusion are: 1) the pitch of the stimuli, 2) the vowels of the stimuli, and 3) the phonetic features of the stop consonants. Let us consider pitch and vowel requisites together. In the case of psychoacoustic fusion the two stimuli must have the same pitch and the same vowel. In the case of phonetic feature fusion, the two stimuli must have either different pitches or different vowels; these differences insure that the phonetic features of the stimuli cannot be acoustically averaged. The other important dimension concerns the features of the to-be-fused stimuli. In psychoacoustic fusion the stop consonants must differ in only the place feature, and the voicing feature must be shared by the two stimuli (for example, /ba/ and /ga/). In phonetic feature fusion, on the other hand, the two stimuli must differ along both dimensions, place and voicing (for example, /ba/ and /ta/).

It is possible that psychoacoustic fusion and phonetic feature fusion may occur at the same time for the same competing stop-vowel stimuli. If /ba/ and /ta/ are presented at the same pitch, an ambiguous experimental situation results. The identical pitches and the shared vowel of the two stimuli set up a situation in which psychoacoustic fusion may occur. The unshared place and voicing features of the two stops, on the other hand, set up a situation in which phonetic feature fusion may occur. If the subject reports hearing /da/, we cannot be sure if the fusion is purely phonetic in nature, or if an element of psychoacoustic averaging contributed to the percept. Perhaps both processes are involved. In such cases, the fusion of /ba/ and /ta/ at the same pitch may be a hybrid of psychoacoustic fusion and phonetic feature fusion.

A note on dichotic competition, perceptual rivalry, and perceptual fusion. Dichotic stop-vowel stimuli are normally thought to "compete" with one another. This competition is perhaps more clearly defined as perceptual rivalry. When two CV syllables are presented dichotically, the subject typically reports hearing one or both of them. The stimuli are rivals, and we have thought that they are not usually combined into a single percept. However, as we have seen in psychoacoustic fusion (fusion type 3) and in phonetic feature fusion (fusion type 4) the subject often fuses stop-vowel stimuli. Thus, perceptual rivalry and perceptual fusion appear to converge, since both processes can occur for the same stimuli.

Although rivalry and fusion may occur simultaneously within the same stimuli, they do not appear to occur at the same level. Consider phonetic feature fusion. Given /ba/-/ta/ the subject never reports hearing /tba/ or /bta/; the phonological constraints of English do not permit two stop consonants

to cluster within a syllable. Thus, we have a case of clear phonological rivalry; at the phoneme level the two stop consonants compete for the same processor. At another level, however, there is fusion. The /b/ and /t/ do not share the same values of place and voicing features. Any combination of labial and alveolar features with voiced and voiceless features yields a permissible stop consonant. Since there are no shared place and voice features in /b/ and /t/, there is no phonetic rivalry between them. The result of this pairing is often fusion at the phonetic level and rivalry at the phonological level.

A similar pattern occurs in psychoacoustic fusion. Given /ba/-/ga/ the subject never reports hearing /bga/ or /gba/; again, there is a clear phonological rivalry. At another level, however, there is fusion, this time at a psychoacoustic level. The place feature of /b/ merges with the place feature of /g/, and the intermediate phoneme /d/ is perceived.

5. CHIRP FUSION: Perceptual construction of phonemes from speech and nonspeech stimuli.[4]

Rand (in preparation) discovered a fifth type of fusion. In chirp fusion there are no competing phonemes, nor is there information loss; instead, different parts of the same speech signal are presented to either ear. The fifth display of Figure 1 shows the syllable /da/ divided into two stimuli. One stimulus contains all the acoustic information in /da/ except the F2 transition, namely the entire first formant and most of the second formant. It is important to note that /da/ without the F2 transition is difficult to identify and is more readily perceived as /ba/ than /da/. The F2 transition alone is the second stimulus. It is very brief--a rapidly falling pitch sweep similar to a bird's twitter, hence the name "chirp." When the /da/ chirp is presented to one ear and the remaining portions of the syllable to the other, the subject reports hearing the full syllable /da/ plus the nonspeech chirp.

Perhaps the most interesting aspect of chirp fusion is that the subject hears more than one auditory image. As in phonetic feature fusion (fusion type 4) he hears two sounds, but he does not hear two speech sounds; instead, he hears one speech sound, /da/, and one nonspeech sound, the chirp. Note that the perceptual whole is greater than the sum of the parts: the subject "hears" the chirp in two different forms at the same time. One form is in the complete syllable /da/, which would sound more like /ba/ without the chirp. The second form is similar to the F2 transition heard in isolation--a nonspeech chirp.

Chirp fusion is more complex than previous fusions we have considered. Pilot studies indicate that many of the energy characteristics of the /da/ chirp may be different from those of the remainder of the /da/ syllable. For example, the two stimuli may have different pitches and fusion rates appear to be unattenuated. Relative differences of 20 Hz appear to have no effect on fusion response levels. Relative intensity differences also do not affect chirp fusion; chirp fusion occurs even if the chirp stimulus is decreased by as much as 30 db relative to the "chirpless" /da/. The chirp in this case is only about

---

[4]The material gathered for this section has come from numerous discussions with Tim Rand.

1/1000 as loud as the "chirpless" /da/. This intensity difference is more impressive when one considers that the same chirp (at -30 db), when electrically mixed with the chirpless /da/ and presented monaurally (instead of dichotically), still sounds more like /ba/ than /da/. Thus, for these stimuli, the chirp is a more potent speech cue when presented to the opposite ear than when presented to the same ear concatenated onto the chirpless stimulus.

Large differences in intensity or pitch between the two stimuli have no apparent effect on fusion rate in chirp fusion. This fact, coupled with "hearing" the chirp in both a speech and nonspeech form suggests that chirp fusion is a more central, higher-level process than the four types of fusion previously discussed. There is yet another dimension which distinguishes it from the lower-level fusions--timing. In chirp fusion the chirp and the chirpless stimulus need not begin at the same time. Rand has done pilot work which indicates that the relative onsets of the two stimuli may differ by as much as $\pm$ 25 msec. In other words, the chirp stimulus can begin 25 msec before the onset of the chirpless /da/, the two stimuli may be simultaneous with respect to their relative onsets, or the chirp can begin 25 msec after the onset of the chirpless /da/. The result for all three cases appears to be the same: the subject hears /da/ plus a chirp. When relative onsets of greater than 25 msec are used, fusion breaks down and the subject begins to hear /ba/ plus a chirp. Nevertheless, relative onset differences tolerated in chirp fusion are much greater than those permissible in the lower-level fusions.

In both phonetic feature fusion (fusion type 4) and chirp fusion (fusion type 5) the features of the two stimuli are combined to form a new speech unit. In phonetic feature fusion, place and voicing information is extracted from separate sources. In chirp fusion, manner and voicing cues in the chirpless stimulus are combined with the place feature extracted from the chirp. Thus, both fusions operate on the level of a single phoneme. By contrast, the next type of fusion to be considered deals with the combination of phonemes into a cluster.

6. PHONOLOGICAL FUSION: Perceptual construction of phoneme clusters.

Day (1968) was first to discover that compatible stop + liquid strings could fuse into one unit: given BANKET and LANKET presented to opposite ears, the subject often hears BLANKET (BANKET/LANKET$\longrightarrow$ BLANKET[5]). One of the unique aspects of phonological fusion is that, unlike psychoacoustic fusion (fusion type 3) and phonetic feature fusion (fusion type 4), two stimuli which contain entirely different segments are presented at the same time, and yet there is no information loss. The segments of both stimuli are combined to form a new percept which is longer and more linguistically complex than either of the two inputs. The sixth display of Figure 1 shows a sample phonological fusion: the inputs /da/ and /ra/ often yield the fusion /dra/. This fusion contains all the linguistic information available in both stimuli. Thus, unlike other fusions, there is no phonological rivalry between the two stimuli.

---

[5]The arrow ($\longrightarrow$) should be read as "yields."

Note that there is a preferred order of phonemes in the fusion response for these stimuli. Given BANKET/LANKET the subject almost never reports LBANKET. Instead, all of his fused responses are BLANKET. This finding appears to be based on the phonological constraints of English: initial liquid + stop clusters are not allowed. Day (1970) has shown that when such constraints are removed, fusion occurs in both directions. Given the stimuli TASS/TACK, subjects give both TASK and TACKS responses.

Day (1968) has shown that phonological fusion is not a function of response biases for acceptable English words. Both a stop stimulus and a liquid stimulus must be present for a stop + liquid fusion to occur. If one presents different productions of BANKET to either ear, the subject reports hearing BANKET (BANKET/BANKET⟶ BANKET). That is, the subject does not report the acceptable English word that corresponds most closely to the nonword inputs. Likewise, LANKET/LANKET⟶ LANKET. However, if the stimuli are BANKET/LANKET, the subject often reports hearing BLANKET regardless of which stimulus is presented to which ear. Furthermore Day (in preparation-a) has shown that even if the subject is informed as to the nature of the stimuli, he still fuses: he perceives BLACK both when the stimuli are BACK/LACK and when they are BLACK/BLACK.

Although there are undoubtably peripheral elements involved in phonological fusion, various studies suggest that it is primarily a central, higher-level cognitive process. Phonological fusion is insensitive to large energy differences between the dichotic stimuli. The stimuli PAY and LAY fuse to PLAY regardless of their pitch. Relative differences of 20 Hz in fundamental frequency have no effect on the characteristics of fusion responses (Cutting, in preparation-a). Intensity differences also have no effect: one stimulus may be 15 db louder than the other and fusion rates do not change (Cutting, in preparation-a).

Phonological fusion is also insensitive to gross differences in the relative onsets of the two stimuli. Day (in preparation-b) has shown that the stimulus onsets may be staggered by as much as 150 msec and fusion rates do not change substantially. Note that these relative onset times are considerably longer than those permissible in any other fusion. Fusion occurs if the stimuli are simultaneous, if the stop stimulus (BANKET) begins 150 msec before the liquid stimulus (LANKET), or even if the liquid begins 150 msec before the stop. Thus, within a wide range of relative onsets, the actual ordering of the phonemes in real time appears to have little effect on fusion rate.

Phonological fusion also appears to be insensitive to changes in dimensions other than pitch, intensity, and timing. Subjects fuse whether the stimuli were uttered by the same vocal tract or by vocal tracts of different sizes (Cutting, in preparation-a). For example, PAY/LAY⟶ PLAY even if PAY has been synthesized to resemble the utterance of a normal adult male, and LAY has been synthesized to resemble the utterance of a midget or small child.

Phonological fusion occurs most readily when the same vowel follows both the initial stop and the initial liquid. Nevertheless, while PAY/LAY⟶ PLAY and GO/LOW⟶ GLOW, the pairing of PAY and LOW can yield PLO or PLAY. Fusion rates are reduced here, but are still at a fairly substantial level (Cutting, in preparation-b). In fact, fusion even occurs with stimuli that have almost no phonemes in common. Day (1968) has shown that two stimuli as different as

BUILDING and LETTER can "fuse" into BILTER, LILTER, or even BLITTERING; such cases involve phonemic exchanges between the two stimuli.

One distinction which appears to be unique to phonological fusion is that not all subjects fuse. In the five types of fusion previously discussed it appears that all subjects fuse equally readily. In phonological fusion, on the other hand, using natural speech stimuli, Day (1969) has shown that some subjects fuse on nearly all trials, while others fuse only occasionally, if at all. Moreover, few subjects score between these extremes. Cutting (in preparation-a), using synthetic speech stimuli, has also found a bimodal distribution of subjects with respect to their fusion rates. Large individual differences are found in many higher-level processes. For example, Turvey (in press) has shown in vision research that individual differences are larger for central masking than for peripheral masking. It appears that in the case of phonological fusion we have a task which is complex enough so that alternative modes of processing are possible. Studies are now underway to explore the possibility that groups of people who perform differently on the fusion task may also retain their group identity on other auditory and visual tasks (Day, in preparation-c).

Finally, phonological fusion appears to have certain linguistic constraints that no other fusion has. For example, some consonant + liquid stimuli fuse more readily than others. Day (1968) has shown that stop + /l/ stimuli fuse more readily than stop + /r/ stimuli. Day (1968) and Cutting and Day (1972) have shown that stop + /r/ stimuli often elicit a stop + /l/ response, whereas the reverse situation rarely occurs. Thus, PAY/RAY may yield PLAY. Cutting (in preparation-a) has found that stop + liquid stimuli fuse more readily than fricative + liquid stimuli: BED/LED⟶ BLED more readily than FED/LED⟶ FLED. These findings cannot be accounted for by the relative frequency of occurrence of these clusters in English. In fact, frequency data show the reverse trends: stop + /r/ clusters outnumber stop + /l/ clusters (Day, 1968) and /f/ + liquid clusters outnumber most other consonant + liquid clusters (Cutting, in preparation-a).

Phonological fusion appears to be the highest level process considered in this paper. Like other higher-level fusions it is insensitive to large stimulus differences in pitch, intensity, and timing. Other dimensions may also be varied with little effect, such as vocal tract size and vowel context. Certain variables, however, do affect fusion rate. Day (1968) has shown that semantics at the word level is one such variable. Fusion occurs most readily when the fused percept is a meaningful word, although nonword fusions do occur (e.g., GORIGIN/LORIGIN⟶ GLORIGIN). Cutting (in preparation-a) has shown that semantics at the sentence level can also influence fusion. Fusion rates are higher when the fusable pair appears in a sentence context. For example, PAY/LAY⟶ PLAY more readily when the stimuli are THE TRUMPETER PAYS FOR US and THE TRUMPETER LAYS FOR US than when PAY and LAY are presented as an isolated pair.

REVIEW OF FUSIONS

We have looked at six phenomena in which stimuli are sent separately to each ear and the subject is asked to report what he heard. The effect of changing various parameters between the two inputs is summarized in Table 1.

TABLE 1: Dimensions which are relevant for the separation of lower- and higher-level fusions. Tolerances are listed within each cell. Specific numbers reflect current knowledge.

	PITCH	INTENSITY	TIMING
LOWER-LEVEL FUSIONS			
1. Sound Localization	< 2 Hz	< 2 db	< 2 msec
2. Spectral Fusion	< 2 Hz	*	< 5 msec
3. Psychoacoustic Fusion	< 2 Hz	*	*
HIGHER-LEVEL FUSIONS			
4. Phonetic Feature Fusion	20 Hz	30 db+	25 msec+
5. Chirp Fusion	20 Hz	30 db	25 msec
6. Phonological Fusion	20 Hz	15 db	150 msec

* systematic data not available
\+ indirect evidence

1. Sound localization occurs for all audible sounds, speech and nonspeech. The first display of Figure 1 shows that when /da/ is presented to both ears at the same time, pitch, and intensity, the subject perceives one /da/ localized "in the center of the head," or at the midline. Pitch variations of 2 Hz can inhibit fusion, such that two stimuli will be heard. Intensity variations of 2 db are sufficient to change the locus of the fusion. Timing differences of 2 msec are sufficient to cause the fused percept to disintegrate into two elements.

2. Spectral fusion occurs for speech sounds and for complex nonspeech sounds. The second display in Figure 1 shows that when F1 of /da/ is presented to one ear and F2 to the other, the subject perceives the fused /da/. Pitch variations of 2 Hz can inhibit fusion. Timing differences of about 5 msec can inhibit the spectral fusion of metronome ticks.

3. Psychoacoustic fusion probably occurs for both speech sounds and nonspeech sounds. We have considered only speech sounds. For example, in the third display of Figure 1, /ba/ is presented to one ear and /ga/ to the other at the same pitch. The resulting perception is /da/. Pitch differences of 2 Hz can inhibit fusion.

4. Phonetic feature fusion occurs for competing speech segments. In the fourth display we note that when /ba/ and /ta/ are presented at different pitches, the subject often reports hearing the "blend" /da/. Pitch does not appear to be an important parameter in this fusion; preliminary work suggests that differences of 20 Hz are easily tolerated and fusion rates are unattenuated. Intensity has not been systematically explored, but data from other sources suggest that differences of as much as 30 db will not inhibit fusion. Our knowledge concerning the effect of timing differences is also indirect. Relative onsets greater than 25 msec inhibit fusion.

5. Chirp fusion is demonstrated in the fifth display of Figure 1. The "chirpless" /da/ is presented to one ear and the /da/ chirp to the other. Subjects perceive the entire syllable /da/ plus a nonspeech chirp. Pilot work has shown that pitch differences of 20 Hz do not alter fusion rates. Intensity differences of 30 db also do not affect chirp fusion. Relative onset differences of 25 msec are tolerable in chirp fusion, but larger relative onset differences inhibit fusion.

6. Phonological fusion occurs for pairs of phonemes which can form clusters, for example BANKET/LANKET→ BLANKET. Display six shows that when /da/ is presented to one ear and /ra/ to the other, the subject often perceives /dra/. Large differences in pitch (20 Hz) and in intensity (15 db) do not appear to affect fusion rate. Gross differences in timing also appear to have no effect; differences of as much as 150 msec in the to-be-fused stimuli do not appear to alter the rate of fusion for stimuli such as BANKET/LANKET. Variations in vocal tract size are also tolerated. Certain linguistic variables, however, do influence fusion rate: two such variables are the types of consonant and liquid phonemes involved in the fusion, and the semantic context in which the stimuli appear.

CONCLUSION

There are at least six different types of fusion in auditory perception. They have been compared and contrasted with respect to three primary parameters: pitch, intensity, and timing. The first three fusions discussed (sound localization, spectral fusion, and psychoacoustic fusion) are sensitive to small changes in any of these parameters. Sensitivity to small differences in stimulus energy and stimulus timing has been noted to be a property of peripheral mechanisms (Turvey, in press). Thus, for the purposes of this paper, these three fusions are considered lower-level, peripheral processes.

In contrast, the other three fusions (phonetic feature fusion, chirp fusion, and phonological fusion) appear to be higher-level processes. In general, these fusions are insensitive to large stimulus differences in pitch, intensity, and timing. Since relative insensitivity to stimulus energy and stimulus timing has been noted as a property of central mechanisms (Turvey, in press) these three fusions may be considered higher-level, central processes.

The six fusions run the gamut from primitive to highly complex levels of processing. Both man and animals can localize sound, a very low level of fusion. Phonological fusion, at the other end of the fusion continuum considered here, appears to be a situation complex enough to allow alternative modes of processing: some subjects fuse the stimuli presented to opposite ears and give a single linguistic response, while others do not. All six processes that we have considered are fusions: the subject combines two signals into a single percept. Yet they are clearly different in many ways. Therefore, we have described, compared, and contrasted the various kinds of fusion so that more precise experimental questions can be posed in order to unravel the processes involved.

REFERENCES

Broadbent, D. E. (1955) A note on binaural fusion. Quart. J. Exp. Psychol. 7, 46-47.

Broadbent, D. E. and P. Ladefoged. (1957) On the fusion of sounds reaching different sense organs. J. Acoust. Soc. Amer. 29, 708-710.

Cutting, J. E. (in preparation-a) The locus of perception in phonological fusion. Ph.D. thesis, Yale University (Psychology).

Cutting, J. E. (in preparation-b) Phonological fusion of "incompatible" stimuli.

Cutting, J. E. and R. S. Day. (1972) Dichotic fusion along an acoustic continuum. J. Acoust. Soc. Amer. 52, 175(A). (Also in Haskins Laboratories Status Report on Speech Research SR-28, 103-114.)

Day, R. S. (1968) Fusion in dichotic listening. Unpublished Ph.D. thesis, Stanford University (Psychology).

Day, R. S. (1969) Temporal order judgments in speech: Are individuals language-bound or stimulus-bound? Paper presented at the 9th annual meeting of the Psychonomic Society, St. Louis. (Also in Haskins Laboratories Status Report on Speech Research SR-21/22, 71-88.)

Day, R. S. (in preparation-a) Dichotic fusion despite knowledge of the stimuli.

Day, R. S. (in preparation-b) Release of language bound perception.

Day, R. S. (in preparation-c) Individual differences in language abilities.

Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. Unpublished Ph.D. thesis, University of Minnesota (Psychology). (Also, Supplement to the Haskins Laboratories Status Report on Speech Research.)

Rand, T. C. (in preparation) Dichotic release from masking for speech.

Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Studdert-Kennedy, M., D. Shankweiler, and D. B. Pisoni. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. Cog. Psychol. 3, 455-466.

Studdert-Kennedy, M., D. Shankweiler, and S. Schulman. (1970) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. Acoust. Soc. Amer. 48, 599-602.

Thompson, C. L., M. R. Stafford, J. K. Cullen, L. F. Hughes, S. S. Lowe-Bell, and C. I. Berlin. (1972) Interaural intensity differences in dichotic speech perception. J. Acoust. Soc. Amer. 52, 174(A).

Turvey, M. T. (in press) On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with pattern stimuli. Psychol. Rev. (Also in Haskins Laboratories Status Report on Speech Research SR-29/30, 1972, 1-92.)

Constructive Theory, Perceptual Systems, and Tacit Knowledge*

M. T. Turvey+
Haskins Laboratories, New Haven

In preparing my comments on Mace's paper[1] I found myself in the pleasant position of having nothing to criticize. I am sympathetic to Gibson's (1966) theory of perception and it seems to me that Mace's comments on Gibson and the constructivist alternative are both justified and well put. I will therefore use this opportunity to touch upon three topics which are related, if only tangentially, to the issues discussed by Mace. First, I will make some additional comments on constructive theory with special reference to the domain of such a theory; second, I will address myself to the idea that "perceptual systems" as defined by Gibson can play several different roles in the perceptual process; and third, I will comment on analysis-by-synthesis for the purpose of drawing a distinction between tacit and explicit identification along the lines suggested by Michael Polanyi (1964, 1966).

Constructive Theory and Linguistic Perception

Constructive theory assumes that perceptual experience is not a direct response to stimulation. Rather, the perceptual experience is constructed or created out of a number of ingredients, only some of which are provided by the sensory stimulation. Other ingredients in a perception recipe are provided by our expectations, our biases, and our knowledge of the world in general.

In view of most students of the constructivist leaning all perceptual experiences are constructed "...from fleeting fragmentary scraps of data signalled by the senses and drawn from the brain's memory banks--themselves constructions from snippets of the past" (Gregory, 1972). The extreme constructivist position expressed in this quote (and criticized by Mace) is conveniently satirized in an analogy drawn by Gilbert Ryle (1949). A prisoner has been held in solitary confinement since birth. His cell has no windows but there are some cracks in the walls through which occasional flickers of light may be seen, and

---

*Presented at the Conference on Cognition and the Symbolic Processes at Pennsylvania State University, October 1972, and to be published in the conference proceedings.

+Also University of Connecticut, Storrs.

[1] Mace, W. M. Ecologically stimulating cognitive psychology: Gibsonian perspectives. (A paper presented at the Conference.)

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

through the stones occasional tappings and scratchings may be heard. On the basis of these snippets of light and sound our prisoner-hero becomes apprised of unobserved happenings outside his cell such as football games, beauty pageants, and the layout of edges and surfaces. In order for our prisoner to perceive these things he must, of course, know something about them in advance. But we should ask how he could ever come to know anything about, say, football games except by having perceived one in the first place?

Ryle's analogy underscores the fact that constructivism in its extreme form takes as its departure point traditional image optics rather than Gibson's (1966) ecological optics; it denies the richness and variety of stimulation at the receptors and consequently denies the elaborateness of the perceptual apparatus. But if we accept Gibson's arguments for information in stimulation and for perceptual machinery capable of detecting that information then the extreme constructivist view is unnecessary. Thus, for example, given Mace's arguments and demonstrations we do not have to interpret Wallach and O'Connell's (1953) kinetic depth effect as a perception synthesized out of information collected over a period of time (cf. Neisser, 1968, 1970). That is to say we do not have to interpret the perceptual experience of a rigid, three-dimensional rotating object as being the result of combining successive retinal snapshots of a two-dimensional form. The constructivist interpretation of the kinetic depth effect arises, in part, from the failure to appreciate that transformations of patterns are probably more stimulating and informative than the static patterns themselves.

The main thrust of Gibson's theory, vis-a-vis constructivism, is that there are complex variables of stimulation which specify directly the properties of the world. Perception of the environment corresponds simply and solely to the detection of these variables of stimulation and there are no intermediary intellectual steps needed to construct perception out of what is detected. Gibson, of course, does not argue that all perception is of this kind; that is, he does not argue that all experiences called perceptual are a direct function of stimulation. Indeed, he admits that some of the experiences called perceptual are not a function of stimulation at all (Gibson, 1959:466). However, he does believe that perception is exclusively a function of stimulation where conditions of stimulation permit.

One apparent exception to Gibson's principle of direct perception is the perception of either the spoken or the written language. Given what we know about speech perception in particular, (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967) and language perception in general, we can state in a paraphrase of Gibson that the conditions of linguistic stimulation do not permit direct perception. It is quite evident that the comprehension of linguistic items received by ear and by eye relies heavily on the context in which the items are occurring. The perception of both spoken and written language proceeds faster than it should and it is remarkably unaffected by a variety of omissions and errors. Thus, our interpretation of a verbal item in normal spoken or written discourse is in some part dependent on our prediction of what the event might be and is not simply dependent on the stimulation provided by the item itself. Our predictions of--or expectations about--a linguistic event derive from three major sources: our knowledge of what has just been perceived; our internal model of the language, i.e., our knowledge of the various linguistic rules; and our knowledge of the world.

Of course, what I have just described characterizes the approach to perception known as analysis-by-synthesis, an approach which has assumed a central role in modern constructivist theory (see Neisser, 1967). Yet while analysis-by-synthesis and constructive theory may prove to be useful to our understanding of the perception of linguistic information by ear and by eye (although it goes without saying that not everybody necessarily agrees; see Corcoran, 1971) they may not prove particularly useful, or even relevant, to other kinds of perception. There are many good reasons for believing that speech perception and reading are rather special perceptual activities and that they may not be representative of how perception occurs in general. To begin with, speech perception appears to involve an articulatory model--i.e., a production model (see Liberman et al., 1967). Both experiment (e.g., Corcoran, 1966; Klapp, 1971) and clinical observation (e.g., Geschwind, 1970) suggest that reading is at least in part parasitic upon the mechanisms of speech. There is no compelling evidence to suggest that other forms of perception proceed by reference to a production model. The special character of linguistic perception is further supported by Mattingly's (1972) argument that grammar emerged as an interface between two mismatched nonlinguistic systems which had evolved separately. On the one hand we have the mechanisms concerned with transmission--the ear and the vocal apparatus--and on the other we have an intellect which represents, rather amorphously I suspect, the world of experience (i.e., the mechanism of long-term memory). Grammatical codes, therefore, convert representations of experience into a form suitable for efficient acoustical transmission, or they convert phonetic events into a form suitable for long-term storage (Liberman, Mattingly, and Turvey, 1972). [And surely this kind of radical conversion is at the heart of constructivist theory; both linguistic perception and linguistic memory are restructurings of stimulation. But we should ask, as Liberman (1972) has, whether such radical conversions occur in other perceptual situations.]

It is perhaps instructive to note that hemispheric damage which results in the reading impairment generally referred to as word-blindness or alexia, may leave unimpaired the ability to name objects (Howes, 1962). But more important to our present concerns is the observation that alexic patients generally have no difficulty perceiving the spatial aspects of things such as distance, shape, size, and movement, that is, the properties of stimulation Gibson is primarily concerned with. We should also note a rather perplexing observation reported by Kohler (1951) concerning the Innsbruck investigations on the reversal of the visual world by means of prisms. After several weeks or months of wearing prisms which reverse the visual world, the visual world may quite suddenly return to normal. But when this reversal of the visual world to normal occurs writing may remain reversed; the perception of written language apparently involves at some level a special visual process. The point is that answers to the question: "how do we perceive linguistically?" should not be viewed as answers to the question: "how do we perceive?"

Perceptual Systems Do More Than Register Invariances

Traditionally the senses have been conceptualized as passive conduits which transmit imperfect images from the retina to the brain where they are represented as collections of raw sensations and out of which perception is eventually fashioned. For each kind of sensory experience there is, reputedly, a special sense; thus, the special sense of vision is the source of visual sensation, the special sense of proprioception is the source of the sensations

of ones own movements, and so on. The convention, of course, has been to classify the senses by modes of conscious quality.

By contrast Gibson proposes that the senses are active systems which register the invariant structures of available stimulation furnished at the receptors and which afford to the observer direct knowledge of his environment. In this view, the "senses"--which Gibson prefers to call perceptual systems--detect information rather than yield sensations and are classified by modes of activity rather than by modes of consciousness. Thus looking and listening replace, respectively, the having of visual and auditory sensations. Of further importance is the idea that a particular kind of information is not necessarily the special domain of a particular perceptual system, but rather that different systems can detect the same information either singly or in combination.

Gibson's substitution of the concept of perceptual systems for that of the senses is a commendable one, and its far-reaching implications for a general theory of perception have been spelled out in the papers by Mace and Shaw[2] in this conference. What I wish to touch upon in the present discussion is the idea that perceptual systems are flexible machineries which can be put to uses other than that of discovering invariants in changing stimulation, although that is their primary function. Thus, in addition to detecting invariances perceptual systems can be generative devices which construct perceptual experiences of certain kinds. But we should guard against concluding that just because perceptual systems <u>can</u> construct, then the everyday perception of the everyday world <u>is</u> constructed. As I view it, relatively few perceptual experiences are constructed. While there is certainly an intimate and theoretically provocative relation between the workings of a perceptual system as a detector of invariances and the workings of a perceptual system as a generative device, I do not think that the relation is one of identity.

There is certainly nothing novel in the idea that a perceptual system can be generative. Indeed, B. F. Skinner (1963) has elegantly expressed this notion in his choice phrase describing the behaviorist position on conscious experience: "seeing does not imply something seen." If I understand Skinner correctly he is saying that seeing is (can be?) a behavior and therefore seeing a Rolls-Royce, for example, is an activity which can be evoked (given the right contingencies) even though no Rolls-Royce is present to be seen. It is instructive to note that the statement which Skinner finds so admirably descriptive of the behaviorist viewpoint is the very kind of statement which expresses the position advanced by constructivists (Gregory, 1972; Kolers, 1968; Neisser, 1967), although I suspect that Skinner and the constructivists find this statement appropriate for different reasons. In any event, the idea that a perceptual system may yield an experience in the absence of stimulation has been well recognized. Thus, dreaming, hallucinating, illusioning, and imaging may all be considered as examples of this characteristic of perceptual systems. But while it is reasonable to propose that a person who is seeing or hearing or smelling things that are not present must be generating them for himself, we need not be convinced by this that the generative mechanisms he uses overlap with the normal mechanisms of seeing, hearing, and smelling. Fortunately there are more solid grounds for inferring the overlap.

---

[2]Shaw, R. Algoristic foundations of cognitive psychology.

In a signal detection experiment a subject is asked to image something and to indicate when he has a good image. At that point a signal is either presented or not and the subject is required to report whether the signal did or did not occur. If the subject was entertaining an auditory image and the signal was auditory then sensitivity (measured as d') is poorer than if a visual image had been entertained. Similarly, the detection of a visual signal is impaired more significantly by concurrent visual imagery than by concurrent auditory imagery (Segal and Fusella, 1970, 1971). The interpretation given to this outcome is that detecting, say, a visual signal and generating a visual image require the services of a common mechanism.

A similar conclusion can be drawn from the work of Brooks (1968). A subject is required to recall (image) a block F that he has recently studied. With the block F in mind he must signal its corners, signaling those at the bottom and top by "yes" those in between by "no" and starting, say, at the bottom right hand corner. The task is far more difficult if he must signal the sequence of yeses and nos by pointing at an array of yeses and nos than if he signals the sequence verbally. The inference is that imaging the block F and pointing at the visually displayed words both depend on the system for seeing. By way of contrast we can ask the subject to learn a short sentence ("A bird in the hand is not in the bush,") and then to go through the sentence mentally indicating each noun by "yes" and every word that is not a noun by "no." In this case, signaling the yeses and nos by pointing is superior to saying the yeses and nos aloud, presumably because the speech imagery required to maintain the sentence and to go through the sentence conflicts with speaking the yeses and nos.

In a similarly motivated experiment (Brooks, 1967) the subject is instructed about the arrangement of digits in a matrix. The subject is told to image the matrix and then to allocate the digits in the matrix according to the instructions which are presented to him either in a written form or aurally. His subsequent recall of the location of the digits in the matrix is poorer in the reading condition than in the listening condition. The inference in this case is that reading a message is antagonistic to the simultaneous representation of spatial relations, whereas listening to a message is not.

Other experiments have pointed to this dependence of memory on the perceptual apparatus relevant to the to-be-remembered material. Thus Atwood (1971) showed that an irrelevant visual perception interfered more with verbal learning by means of imagery than did an irrelevant auditory task. Den Heyer and Barret (1971) showed that the short-term retention of the digits in a matrix was interfered with more by a verbal interpolated task than by a visual interpolated task, while the reverse was true for the retention of the spatial location of the digits.

On this evidence we may conclude that perceiving and imaging engage the same neural apparatus, at least at some level, and that memory sustaining operations (such as rehearsal) and acts of remembering (such as imaging) are carried out within the perceptual system most related to the memory material. In other words, there is support for the argument that a perceptual system is also a generative system.

It is commonplace to regard imagery and hallucinating experiences and the like as the arousal of stored representations. The use of nouns such as "image, hallucination, dream," etc., commits us to the idea of something--an object or a scene--which is recalled or rearoused or constructed and which then--like a real object or scene--is viewed or experienced by an observer. Alternatively we could argue that it is the act of imaging or dreaming or hallucinating that is experienced, and that (following Skinner, 1963) imaging, dreaming, and hallucinating do not imply things imaged, dreamt, or hallucinated. On this argument, which I happen to prefer, it is true to say that I am imaging my grandmother, but it is not true to say that I have an image of my grandmother in my head.

Related to the generative capability of perceptual systems is a rather important use for at least one perceptual system, the visual perceptual system: to model or imitate external events. In a sense, all acts of construction carried out by a perceptual system are imitative acts, but what I have in mind here is the idea of the visual perceptual system functioning as an analog spatial model in which orderly, physical operations can be conducted vicariously (cf. Attneave, 1972). This characterization of the visual system is similar to Craik's (1943) thesis that the brain is essentially a complex machine which can parallel or model physical processes, a capability of neural machinery which Craik views as the fundamental feature of thought and explanation.

Evidence that the visual perceptual system can model physical space, i.e., that it can exhibit processes which have a similar relation structure to physical space (cf. Craik, 1943) is to be found in experiments conducted by Shepard and his colleagues. In one experiment (Shepard and Metzler, 1971) subjects were shown a pair of two-dimensional portrayals of three-dimensional objects and were asked to decide as quickly as possible whether one of the objects could be rotated into the other. The decision latency was shown to be an increasing linear function of the angular difference in the portrayed orientation of the two objects. At 0^{o} difference the latency was 1 sec while at 180^{o} difference the latency was 4 or 5 sec. Each additional degree of rotation added approximately 16 msec to the latency of recognition and this was essentially so whether the rotation was in the plane of the picture or in depth. In a further experiment (Shepard and Feng, 1972) subjects were given a picture of one of the patterns of six connected squares which is produced when the faces of a cube are unfolded and laid out flat. Their task was to decide with minimal delay whether two marked edges of two different squares would meet if the square was folded back into the cube. The time to reach a decision increased linearly with the sum of the number of squares that would bave been involved if the folding up operation were actually performed.

In these experiments of Shepard's the subject is apparently imitating covertly those operations which he would perform if he were actually to rotate a physical object or actually to fold up a physical pattern of squares into a cube. Moreover, these covert motor activities parallel actual motor activities in that they are performed in a continuous space and in real time. On the evidence we should argue that the neural spatial representation which is afforded by the visual perceptual system and in which these covert performances occur is a model of, or an analogue of, physical space.

From other experiments we can infer an interesting complicity between other perceptual systems and the visual one where spatial properties are involved. Auditory localization (Warren, 1970) has been shown to be better with the eyes open than with the eyes closed; learning responses to tactile stimuli delivered in fixed locations is better with unrestricted than with restricted vision (Attneave and Benson, 1969); and the short-term retention of a spatial arrangement of tactile stimulation is impaired significantly more by an irrelevant arithmetic task presented visually than by that same task presented auditorily (Sullivan, in preparation). What these experiments imply is that information about location is mapped into the spatial analogue system provided by vision even when the location information is received or detected by other perceptual systems.

Knowing About Things You Do Not Know You Know About

As I have commented above, analysis-by-synthesis and constructive theory have much in common. The idea of synthesis is a slippery one but we can come to terms with it if we consider the way in which a blindfolded man might attempt to recognize a solid triangular figure by moving his finger around the outline. (The example is taken from an early discussion of synthesis by Mackay, 1963.) To our blindfolded man the concept of triangularity is defined by and symmetrical with the sequence of elementary responses necessary in the act of replicating the outline of a triangle. Now we may presume that the recognition of any sensory event is in some sense an act of replication of the stimuli received. In other words, replicas of the input are generated until there is a significant degree of resemblance between a synthetic replica and the input. Of course the input which the replicating or synthesizing mechanism is dealing with is in quite a different physical form from the original input to the sensory receptors. It is probably in the form of neuroelectrical activity of some spatial-temporal specificity. In any event, to identify a triangle I do not have to synthesize triangles; to identify a smell I do not have to synthesize odors.

Generally speaking, analysis-by-synthesis models propose that identification lies in the act of achieving a reasonable facsimile of the input. But the constructivist view of perception, at least on my understanding of it, may wish to ascribe something more than "identification" to the replicative act. The stronger and preferred position is that the perceptual experience of something corresponds to the act of synthesizing that something. Thus, for example, with reference to the spontaneous reversal of perspective during "midflight" of a Necker cube set into oscillating apparent motion, Neisser (1967:144) comments: "...the reversal of perspective at that point emphasizes that figural synthesis is not a matter of cold-blooded inference but of genuine construction." The experiences of dreaming, hallucinating, and imaging are especially relevant; as I noted earlier, it seems reasonable to propose that a person who is seeing or hearing things that are not present is experiencing his own internal acts of synthesis. But on the constructivist view one wants to argue, in addition, that the perception of an actual event corresponds to an act of synthesis and this in my opinion raises a serious, and as far as I know, unanswered question. Is the act of synthesis which underlies the imaging of, say, a capital A or a loved one's face, the same kind of operation as that which underlies the identification of a capital A or a loved one when they are visually present?

I am aware of very little information which bears on this question. There are, however, a few hints from case studies of agnosia which suggest that the two operations I have referred to are not of the same kind. A patient who cannot read letters, i.e., cannot identify them when they are presented visually, may still be able to visualize them and describe their features. Conversely, a patient who can read letters may not be able to image them at all (Nielsen, 1962:35-40).

Let us hold this somewhat isolated observation in abeyance for a moment and turn to a more serious but related problem. If identification occurs in conjunction with the synthesizing of a reasonable match, and if perceptual experience corresponds to that successful act of synthesis then we should conclude as follows: the conscious perceptual experience of a sensory event is the earliest stage in the processing of that event at which the identification of that event can be said to have occurred. I intend to argue that this conclusion is false and that at least for certain kinds of linguistic material identification precedes the conscious experience and on occasion can be shown to occur in the absence of any conscious experience whatsoever. If my argument is correct then we should suppose that the processes underlying identification and those underlying conscious experience are quite different. To put this another way, the operations by which identification of a capital A (using our earlier example) proceeds and those by which the conscious experience of a capital A is expressed are not identical. Thus we should not be surprised to find, on occasion, brain-injured patients who cannot identify letters but can easily image them.

Michael Polanyi (1964, 1966) has for some time argued for distinguishing between two species of knowledge: tacit knowledge, about which we cannot speak, and explicit knowledge, about which we can. This distinction--adopted here in a rather diluted form--will prove fruitful to the ensuing discussion. I will attempt to show that we may know the identity of a verbal event tacitly, but a further operation--different from that underlying tacit identification--is needed if we are to know the identify of the event explicitly.

A good starting point is provided by the situation evident in visual masking. As you probably know, when two stimuli are presented to an observer in rapid succession perceptual impairment may result. Either the first or the second stimulus may be phenomenally obscured, or at least, not identifiable. One general principle of masking is especially relevant: when masking is of central origin (under conditions of dichoptic stimulation) the later-arriving stimulus is the one likely to be identified rather than the leading stimulus. In short, masking of central origin is primarily backward and this I propose is an important comment on the nature of central processes (Turvey, in press). We should also note that whether or not a lagging stimulus can centrally mask a leading stimulus is dependent on there being some geometric (and/or perhaps semantic) similarity between the two. By way of contrast, masking of peripheral origin can occur in the absence of any formal similarity; in the peripheral domain the comparative energies of the two stimuli are more important (see Turvey, in press).

Paul Kolers (1968) offers a useful analogy for backward masking of central origin. The idea is that the central processor may be likened to a clerk who

receives customers on an aperiodic schedule. When a customer enters the store the clerk asks him a variety of questions in order to determine the customer's dispositions and wants. However, if a second customer enters soon after the first the clerk may be hurried and, therefore, less thorough in his treatment of the first. Consequently, some things may be left undone. But we should note that the clerk has registered and responded to some of the first customer's requests. The analogy emphasizes that although processing of the first stimulus in a backward masking situation may not be completed and consequently an explicit account of the first may not be forthcoming, something about the first may well be known.

One kind of experiment in particular is a rather elegant demonstration of this point. We know that when a stimulus is decreased in physical energy, reaction time to its onset is increased proportionately. However, the reaction time to a backwardly masked stimulus, which may appear either phenomenally decreased in brightness or absent altogether is not so affected (Fehrer and Raab, 1962; Harrison and Fox, 1966; Schiller and Smith, 1966). Thus, we should suppose that in the presence of the masker those operations which determine the phenomenal appearance of the stimulus have been left relatively undone, but those which detect the occurrence of the stimulus and determine its intensity have been completed.

But other experiments are more relevant to the distinction that I seek to draw between tacit and explicit identification. First is an experiment reported by Wickens (1972) which shows that an observer may have some knowledge of the meaning of a masked word even though he might be unable to report the actual identity of the word. In this experiment a word was briefly exposed and followed by a patterned mask. Then the subject was given one of two possible words and asked to guess whether it was similar in some way to the masked and nonidentified word. This second word was never identical to the masked word but it was, half of the time, similar on some dimension to the masked word. The other half of the time it was dissimilar. For some dimensions at least--the semantic differential, taxonomic categories, and synonymity--the subject was likely (better than chance) to identify the semantically related word. The conclusion we may draw from this experiment of Wickens is that one can have tacit knowledge about the meaning of a word in advance of explicit knowledge about its identity. This is also the conclusion I think we should draw from the experiments of Reicher (1969) and Wheeler (1970). Those experiments showed that under identical conditions of backward masking, with careful controls for response-bias effects, a letter could be more accurately recognized if it was part of a word than if it was part of a nonword, or presented singly (cf. Smith and Haviland, 1972). It has always seemed to me that the simplest interpretation of this result is that meaningfulness (and/or familiarity) affects the time taken to process (cf. Eichelman, 1970). But if this is true then we are faced with trying to understand how meaningfulness or familiarity can assist speed and accuracy of identification since we should argue, on the conventional view, that sensory data have to make contact with long-term storage, i.e., have to be identified, before their meaning or familiarity can be ascertained.

This issue is similarly exposed in those experiments which demonstrate a direct relation between the number of syllables or pronounceable units in a verbal event and the time taken to identify it. Thus, for example, Klapp (1971) has shown that the time taken to press a key to indicate that a pair of two-

syllable numbers, e.g., 15 and 15, or 80 and 80, were the same was measurably shorter than the time needed to indicate the sameness of a pair of three-syllable numbers, e.g., 28 and 28, or 70 and 70. The question we should ask of this startling result is: how can the number of syllables affect the time to identify, since surely one must first identify an optical pattern such as 15 or 70 before one can know how to pronounce it?

In a similar vein there is evidence that the category, letter or digit, to which a character belongs can be known before its identity is determined (Brand, 1971; Ingling, 1972; Posner, 1970). In short, we can know that a character is a letter or a digit before we know which letter or digit it is. On Ingling's (1972) data in particular we should have to argue that determining category membership is not based on any simple or obvious feature analysis. In passing we should also note that these demonstrations are in concert with the special cases of visual alexia reported by Dejerine (1892) and Ettlinger (1967). Here injury to the left hemisphere results in an inability to read letters but leaves unimpaired the ability to read arabic numerals. And it is not that the patient has necessarily forgotten the names because he might be able to identify letters conveyed to him tactually. Nor is his problem that of being unable to discriminate letter features since he can sort letters into groups where each group represents one particular letter.

By way of summary, there is good reason to propose that with respect to certain events one can be said to know something about the identity of an event before one knows that event's identity. This seeming paradox, alluded to elsewhere by Coltheart (1972a, 1972b), can be resolved if we distinguish between tacit and explicit identification and view the latter as preceded by and shaped by the former. An experiment by Worthington (1964) shows that one can have tacit knowledge of the semantic character of an event in the absence of any awareness, i.e., explicit knowledge, of its presence. On the surface at least, Worthington's experiment had to do with the time course of dark adaptation. Light adapted subjects seated in a black room were requested to view a designated area in which would appear a dim white light. Their task was simply that of pressing a button as soon as they saw anything in the specified area. Pressing the button turned the light off and the dependent measure was the time elapsed before the button was pressed. Unbeknown to the subjects, the dim light was a disc with a word printed on it in black. The word could be either an obscene word or a geometrically similar neutral word. Worthington found that the average button-pressing latency was determined by the semantic status of the word, with the obscene words yielding longer elapsed times. It is important to note that no subject ever reported seeing anything in the white light.

Further support for the tacit/explicit distinction is to be found in the literature on selective attention in audition, particularly in two experiments. Both use the technique of dichotic stimulation with the shadowing of one of the two concurrent messages. The general finding with this paradigm is that the subject knows little about the unattended message. But I should choose my terms more carefully; the general finding is that the subject knows very little explicitly about the unattended message. At all events, as Cherry (1953) initially observed and as many have confirmed since (e.g., Triesman and Geffen, 1967) a subject may be able to give a relatively detailed account of the physical character of the unattended message but may be sorely limited in his

ability to report on the semantic content of the message. We shall see, however, that the subject knows a great deal more about the unattended message than he can tell.

In one experiment (Lewis, 1970) pairs of words were presented simultaneously such that the unattended message words were associatively related, semantically related, or unrelated to their partners in the shadowed message. Although the subjects were unable to report the words on the unattended channel, it was shown that shadowing reaction time was slower when the word presented in the nonattended message was synonymous with its pair on the shadowed ear. In short, the unattended words were identified but their identification apparently was not made explicit. Similar evidence is provided in a recent experiment by Corteen and Wood (1972). In this experiment certain words were first associated with shock to establish skin-conductance change to these words alone. The shock-associated words were then embedded in the unattended message along with words from the same class (cities) as the shock-associated words, and with control words. Both the shock-associated and nonshock-associated city names produced a significant number of autonomic responses even though the subjects (according to the criteria of awareness employed) were not aware of them.

We should suppose, as I did earlier, that there are important distinctions to be drawn between the processes by which we tacitly know and those by which we explicitly know. To begin with, I suspect that the operations of tacit and explicit identification differ in that the former, unlike the latter, do not make demands on our limited processing capacity. Support for this idea can be drawn from several sources: recent experimental and theoretical analyses of attentional components (Posner and Boies, 1971), attempts to determine the locus of the Stroop effect (Keele, 1972; Hintzman, Carre, Eskridge, Owens, Shaff, and Sparks, 1972), and investigations into the relation between central processing capacity and iconic memory (Doost and Turvey, 1971). Essentially these sources hold that selective attention and limited capacity effects operate after a sensory event has made contact with long-term store (cf. Norman, 1968; Posner and Warren, 1972).

The argument has been made that certain variables which affect identification, such as meaning and familiarity, can only influence the course of perception after contact with long-term store. Thus, in an experiment such as Klapp's (1971) contact between an optical pattern, say, "17," and long-term store, must precede the determination of how that pattern is to be pronounced. Therefore, it must be argued that the number of syllables in the verbalization of the pattern cannot affect the course of tacit identification. On the contrary, the number of syllables can only affect the temporal course of explicit identification. By the same token, it is the conversion from tacit to explicit identification rather than the process of contacting long-term store which is sensitive to meaning and familiarity.

A nonlinguistic analog of the Reicher-Wheeler phenomenon has been reported by Biederman (1972). Essentially, the experiment showed that an object was more accurately identified when part of a briefly exposed real-world scene than when it was part of a jumbled version of that scene, exposed equally briefly. And this was true even when the subject was instructed, prior to exposure, where to look and what to look for. Biederman's discovery implies that the coherency and symmetry of the real-world scene affected the explicit

identification of the particulars of its composition. In a somewhat related experiment Eichelman (1970) has shown that a physical match (see Posner and Mitchell, 1967) is made faster between two words than between two nonwords (cf. Kreuger, 1970).

The question we should ask of all these experiments is: how can "higher order" properties of stimulation, such as symmetry, familiarity, and meaning, affect the identification of the "lower order" properties from which the "higher order" properties are apparently derived? On the present view, the answer to this question is that these higher-order properties are detected by relatively direct means (analogous, perhaps to Gibson's idea of "resonance"), and that explicit knowledge about the particulars, and other kinds of information embodied in the stimulation, is accessible only after such tacit identification.

In sum, pattern recognition can be said to consist of two rather broadly defined stages. The first is that in which stimulation contacts long-term store and the second is that in which the tacit identification afforded by the first stage is converted into explicit knowledge. It would appear on the evidence that the processes involved in the two stages are quite different. Moreover, it would appear that much of what we know about "pattern recognition" is related to the class of operations by which things come out of long-term store, i.e., the tacit-to-explicit conversion, rather than to the manner in which patterns of stimulation contact long-term store in the first place. In short, the "Hoffding Step" (Hoffding, 1891; Neisser, 1967) remains very much a mystery.

In view of the foregoing we might also speculate that the form of knowledge at the tacit level differs from that at the explicit level. This is, of course, the essence of Polanyi's (1964, 1966) argument. Here we should take it to mean that the explicit account of an event and the tacit account of that same event may look quite different, even radically so. Consider if you will the phenomenon in the short-term memory literature known as release from proactive interference (PI). On successive short-term memory tests of the distractor kind (Brown, 1958; Peterson and Peterson, 1959) a subject is given short lists of maybe three words to retain, a new list for each test. If the words presented on the successive tests are drawn from the same category recall performance across the successive tests will decline precipitously. If we now present words on a short-term memory test which have been drawn from a category conceptually different from that used in the immediately preceding tests then there is an abrupt recovery in recall performance. For example, if a subject received three successive tests with digits as the to-be-remembered material and then on the fourth test he was given letters to retain, performance on the fourth test would be equivalent to that on the first and substantially superior to that on the third. Wickens (1970) has proposed that the PI release procedure identifies "psychological" categories. We can assume that there is a common way of encoding within a class (accounting for the decline in recall) which differs between classes (accounting in turn for the increase in recall with shift in class).

Table 1 shows two distant classes of material as defined by PI release. The set of words in the left column consists of a random arrangement of three words drawn from the evaluative dimension, three words from the potency dimension and three words from the activity dimension of the semantic differential (Osgood, Suci, and Tannenbaum, 1957). Each word rates high on one dimension

TABLE 1

farm	wife
prevent	burn
uncle	silence
sea	debt
car	sing
play	young
religious	disease
action	alone
develop	serious

and is relatively neutral on the other two. The right column of words is similarly constructed. The difference between the two columns is that the left-hand column words are drawn from the positive pole of their respective dimensions and the right-hand column words are drawn from the negative pole of their respective dimensions (all words were selected from Heisse, 1965). The experimental evidence is that shifting across dimensions within the same polarity does not yield a release from PI; on the other hand, a highly significant improvement in recall occurs following a change in polarity either within or between dimensions (Turvey and Fertig, 1970; Turvey, Fertig, and Kravetz, 1969). In brief, it has been shown that the polarities are orthogonal but the dimensions are not. What I should like to argue is that this distinction between positive and negative polarity is made only tacitly. In the PI release situation a distinction is obviously being made, and without effort, between the two polarities. But I submit that close examination of Table 1 and careful perusal of the individual words will not lead you to conclude that the two columns differ in any sensible way. Imagine if the words in the two columns were simply mixed together and you were ignorant of the semantic differential (as were the subjects in the experiments). I doubt if you could even begin to sort them into the two categories I have described.

In other words, you can make a distinction tacitly that you cannot readily make explicitly. Quite to the contrary is the situation with nouns and verbs. A shift from nouns to verbs or vice versa does not lead to a release from PI (Wickens, 1970), but one can with some facility distinguish nouns from verbs if one is asked to do so. In the Lewis (1970) experiment referred to above, synonymity between attended and unattended words exerted a marked effect on the reaction time to attended words, but associative relations based on associative norms did not. We might argue from this result that associative norms reflect explicit distinctions but are themselves not isomorphic with the structure of tacit knowledge. Similarly, we can argue that the structure of tacit knowledge does not incorporate images. On the evidence, a distinction is not made tacitly between high-imagery concrete words and low-imagery abstract words, although such a distinction is clearly made explicitly. Wickens and Engle (1970) failed to find PI release with a shift from concrete to abstract words, and vice versa, even though the imagery variable is known to be important in free-recall and paired-associate learning (Paivio, 1969). Imaging, we might suppose, is constructing from tacit knowledge.

Assuming, therefore, that my interpretation of the PI release situation is not too far off the mark, we may draw the following, highly speculative but

intriguing conclusion: you may make distinctions tacitly that you cannot make explicitly, and, conversely, you may make distinctions explicitly that are not furnished tacitly. In this latter case we should assume that such explicit distinctions are constructed.

REFERENCES

Attneave, F. (1972) Representation of physical space. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington: Winston).

Attneave, F. and B. Benson. (1969) Spatial coding of tactual stimulation. J. Exp. Psychol. 81, 216-222.

Atwood, G. (1971) An experimental study of visual imagination and memory. Cog. Psychol. 2, 239-289.

Biederman, I. (1972) Perceiving real-world scenes. Science 177, 77-79.

Brand, J. (1971) Classification without identification in visual search. Quart. J. Exp. Psychol. 23, 178-186.

Brooks, L. R. (1967) The suppression of visualization by reading. Quart. J Exp. Psychol. 19, 289-299.

Brooks, L. R. (1968) Spatial and verbal components of the act of recall. Canad. J. Psychol. 22, 349-368.

Brown, J. (1958) Some tests of the decay theory of immediate memory. Quart. J. Exp. Psychol. 10, 12-21.

Cherry, E. C. (1953) Some experiments on the recognition of speech with one and with two ears. J. Acoust. Soc. Amer. 25, 975-979.

Coltheart, M. (1972a) Visual information processing. In New Horizons in Psychology, ed. by P. C. Dodwell. (Harmondsworth: Penquin Books).

Coltheart, M. (1972b) Readings in Cognitive Psychology. (Toronto: Holt, Rhinehart, Winston).

Corcoran, D. W. J. (1966) An acoustic factor in letter cancellation. Nature 210, 658.

Corcoran, D. W. J. (1971) Pattern Recognition. (Harmondsworth: Penguin Books).

Corteen, R. S. and B. Wood. (1972) Autonomic responses to shock-associated words in an unattended channel. J. Exp. Psychol. 94, 308-313.

Craik, K. (1943) The Nature of Explanation. (Cambridge: Cambridge University).

den Heyer, K. and B. Barrett. (1971) Selective loss of visual and verbal information in STM by means of visual and verbal interpolated tasks. Psychon. Sci. 25, 100-102.

Dejerine, J. (1892) Contribution a l'étude anatamo-pathologique et clinique des differentes variétés de cécité verbale. Mémoires de la Société de Biologie 4, 61.

Doost, R. and M. T. Turvey. (1971) Iconic memory and central processing capacity. Percep. Psychophys. 9, 269-274.

Eichelman, W. H. (1970) Familiarity effects in the simultaneous matching task. J. Exp. Psychol. 86, 275-282.

Ettlinger, G. (1967) Visual alexia. In Models for the Perception of Speech and Visual Form, ed. by W. Wathen-Dunn. (Cambridge: MIT Press).

Fehrer, E. and D. Raab. (1962) Reaction time to stimuli masked by metacontrast. J. Exp. Psychol. 63, 143-147.

Geschwind, N. (1970) The organization of language and the brain. Science 170, 940-944.

Gibson, J. J. (1959) Perception as a function of stimulation. In Psychology: A Study of a Science, Vol. 1, ed. by S. Koch. (New York: McGraw-Hill).

Gibson, J. J. (1966) The Senses Considered as Perceptual Systems. (Boston: Houghton Mifflin).

Gregory, R. (1972) Seeing as thinking: An active theory of perception. Times Literary Supplement (London) June 23, 707-708.
Harrison, K. and R. Fox. (1966) Replication of reaction time to stimuli masked by metacontrast. J. Exp. Psychol. 71, 162-163.
Heisse, D. R. (1965) Semantic differential profiles for 1000 most frequent English words. Psychol. Monogr. 79, (Whole No. 601).
Hintzman, D. L., F. A. Carre, V. L. Eskridge, A. M. Owens, S. S. Shaff and M. E. Sparks. (1972) "Stroop" effect: Input or output phenomenon? J. Exp. Psychol. 95, 458-459.
Hoffding, H. (1891) Outlines of Psychology. (New York: Macmillan).
Howes, D. (1962) An approach to the quantitative analysis of word blindness. In Reading Disability: Progress and Research Needs in Dyslexia, ed. by J. Money. (Baltimore: Johns Hopkins Press).
Ingling, N. (1972) Categorization: A mechanism for rapid information processing. J. Exp. Psychol. 94, 239-243.
Keele, S. W. (1972) Attention demands of memory retrieval. J. Exp. Psychol. 93, 245-248.
Klapp, S. T. (1971) Implicit speech inferred from response latencies in same-different decisions. J. Exp. Psychol. 91, 262-267.
Kohler, I. (1951) Uber und Wandlurgen der Wahrnehmungswelt. (Vienna: Rudolph M. Rohrer). (Translated by H. Fiss, The formation and transformation of the perceptual world. Psychol. Iss. 3, No. 4.)
Kolers, P. A. (1968) Some psychological aspects of pattern recognition. In Recognizing Patterns, ed. by P. A. Kolers and M. Eden. (Boston: MIT Press).
Kreuger, L. E. (1970) Visual comparison in a redundant visual display. Cog. Psychol. 1, 341-357.
Lewis, J. L. (1970) Semantic processing of unattended messages during dichotic listening. J. Exp. Psychol. 85, 225-228.
Liberman, A. M. (1972) The specialization of the language hemisphere. Paper presented at the Intensive Study Program of the Neurosciences Research Project, Boulder, Colo. [Also in Haskins Laboratories Status Report on Speech Research SR-31/32 (this issue).]
Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington: Winston).
Mackay, D. (1963) Mindlike behavior in artifacts. In The Modeling of Mind: Computers and Intelligence, ed. by K. M. Sagre and F. J. C. Crosson. (New York: Simon and Schuster).
Mattingly, I. G. (1972) Speech cues and sign stimuli. Amer. Scient. 60, 327-337.
Neisser, V. (1967) Cognitive Psychology. (New York: Appleton-Century Crofts).
Neisser, V. (1968) The processes of vision. Sci. Amer. 219, 204-214.
Neisser, V. (1970) Visual imagery as process and as experience. In Cognition and Affect, ed. by J. S. Antrobus. (Boston: Little, Brown).
Nielsen, J. M. (1962) Agnosia, Apraxia, Aphasia: Their Value in Cerebral Localization. (New York: Hafner).
Norman, D. A. (1968) Toward a theory of memory and attention. Psychol. Rev. 75, 722-736.
Osgood, C. E., G. J. Suci, and P. H. Tannenbaum. (1957) The Measurement of Meaning. (Urbana: University of Illinois Press).

Paivio, A. (1969) Mental imagery in associative learning and memory. Psychol. Rev. 76, 241-263.
Peterson, L. R. and M. J. Peterson. (1959) Short-term retention of individual verbal items. J. Exp. Psychol. 58, 193-198.
Polanyi, M. (1964) Personal Knowledge: Towards a Post-Critical Philosophy. (New York: Harper).
Polani, M. (1966) The Tacit Dimension. (Garden City: Doubleday).
Posner, M. I. (1970) On the relationship between letter names and superordinate categories. Quart. J. Exp. Psychol. 22, 279-287.
Posner, M. I. and S. J. Boies. (1971) Components of attention. Psychol. Rev. 78, 391-408.
Posner, M. I. and R. F. Mitchell. (1967) Chronometric analysis of classification. Psychol. Rev. 74, 392-409.
Posner, M. I. and E. Warren. (1972) Traces, concepts and conscious constructions. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington: Winston).
Reicher, G. M. (1969) Perceptual recognition as a function of stimulus material. J. Exp. Psychol. 81, 275-280.
Ryle, G. (1949) The Concept of Mind. (London: Hutchinson).
Schiller, P. M. and M. C. Smith. (1966) Detection in metacontrast. J. Exp. Psychol. 71, 32-39.
Segal, S. J. and V. Fusella. (1970) Influence of imaged pictures and sound on detection of visual and auditory signals. J. Exp. Psychol. 83, 458-464.
Segal, S. J. and V. Fusella. (1971) Effect of images in six sense modalities on detection of visual signal from noise. Psychon. Sci. 24, 55-56.
Shepard, R. N. and J. Metzler. (1971) Mental rotation of three-dimensional objects. Science 171, 701-703.
Shepard, R. N. and C. Feng. (1972) A chronometric study of mental paper folding. Cog. Psychol. 3, 228-243.
Skinner, B. F. (1963) Behaviorism at fifty. Science 140, 951-958.
Smith, E. E. and S. E. Haviland. (1972) Why words are perceived more accurately than nonwords: Inference versus unitization. J. Exp. Psychol. 92, 59-64.
Sullivan, E. V. (in preparation) On the short-term retention of serial tactile stimuli. Unpublished Masters thesis, University of Connecticut, Storrs.
Triesman, A. and G. Geffen. (1967) Selective attention: Perception or response? Quart. J. Exp. Psychol. 19, 1-17.
Turvey, M. T. (in press) On peripheral and central processes in vision: Inferences from an information processing analysis of masking with patterned stimuli. Psychol. Rev.
Turvey, M. T. and J. Fertig. (1970) Polarity on the semantic differential and release from proactive interference in short-term memory. J. Verb. Learn. Verb. Behav. 9, 439-443.
Turvey, M. T., J. Fertig, and S. Kravetz. (1969) Connotative classification and proactive interference in short-term memory. Psychon. Sci. 16, 223-224.
Wallach, H. and D. N. O'Connell. (1953) The kinetic depth effect. J. Exp. Psychol. 45, 205-207.
Warren, D. (1970) Intermodality interactions in spatial localization. Cog. Psychol. 2, 114-133.
Wickens, D. D. (1970) Encoding categories of words: An empirical approach to meaning. Psychol. Rev. 77, 1-15.
Wickens, D. D. (1970) Characteristics of word encoding. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington: Winston).

Wickens, D. D. and R. W. Engle. (1970) Imagery and abstractness in short-term memory. J. Exp. Psychol. 84, 268-272.
Wheeler, D. D. (1970) Processes in word recognition. Cog. Psychol. 1, 59-85.
Worthington, A. G. (1964) Differential rates of dark adaptation to "taboo" and "neutral" stimuli. Canad. J. Psychol. 18, 757-768.

Hemiretinae and Nonmonotonic Masking Functions with Overlapping Stimuli

Claire Farley Michaels[+] and M. T. Turvey[++]
Haskins Laboratories, New Haven

Single letter targets followed at varying onset-onset intervals by a patterned mask were presented for identification to the hemiretinae of both eyes. The target and mask stimuli were spatially overlapping; the mask could impede target perception dichoptically and the energy of the target stimuli was twice that of the mask. Under these conditions U-shaped monoptic masking functions were obtained which did not differ, as a function of hemiretina, in their overall shape or in their points of maximal masking.

Recent evidence indicates that U-shaped masking functions are not limited to conditions of metacontrast. Nonmonotonic functions relating degree of masking to stimulus-onset-asynchrony (SOA) for spatially overlapping targets and masks have been reported by Purcell and Stewart (1970), Weisstein (1971), and Turvey (in press). Turvey (in press) has hypothesized that nonmetacontrast U-shaped functions should occur under the following conditions: when the energy of the target is greater than that of the mask, and when the mask can effectively impede the perception of the target under conditions of dichoptic presentation, i.e., the mask is an effective central mask.

An explanation of the U-shaped function obtained when these conditions prevail can be stated quite generally in terms of a gradual shift with increasing SOA from masking of peripheral origin to masking of central origin (Turvey, in press). It is proposed that at zero and at very brief SOAs the induced perceptual impairment is of peripheral origin. At brief intervals the two stimuli, target and mask, engage common peripheral networks and under conditions of peripheral interaction the stimulus of greater energy dominates. Thus, peripherally, a greater energy target will occlude a lower energy mask. At comparatively larger SOAs it is proposed that the two stimuli do not interact peripherally but arrive centrally as separate events. The nature of central processing is such that given the reception of two stimuli in close succession the operations on the earlier stimulus are either terminated or distorted by the arrival of the later stimulus. Centrally the energy relation between the two stimuli is relatively unimportant; what matters is the order of arrival, with the advantage accruing to the later stimulus. Thus, centrally, the later-arriving mask can impede the perception of the greater-energy target. With further increments in SOA the perceptual impairment of the target induced centrally by the after-coming mask declines because more time is allowed for the central processor to determine the target stimulus before the mask arrives.

[+]Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

These notions have received some support in a recent series of experiments reported by Turvey (in press). The present experiment was conducted as a further demonstration of nonmetacontrast U-shaped masking functions under the conditions described above. In addition, the experiment asked whether the overall shape and/or peak of the functions varied with the hemiretina to which the target and mask were presented.

Method. Two three channel tachistoscopes (Scientific Prototype, Model GB) modified for dichoptic viewing were used to present single letter stimuli (A, H, M, T, U, V, W, X, Y) to one of the four hemiretinae. The viewing field of the tachistoscope subtended 6.5 horizontal by 3.5 vertical degrees of visual angle. The lines composing the letters were .15 degrees thick, while the letters were .92 degrees high and, on average, .60 degrees wide. The center of the letters was 1.15 degrees of visual angle to the left or right of a centrally located point of fixation. The target letter was followed monoptically by a pattern mask which consisted of two identical composites of letter fragments, the centers of which were positioned 1.15 degrees to the left and 1.15 degrees to the right of the fixation point. The mask and a target letter are presented in Figure 1. Pilot data had indicated that the mask could successfully impair target identifi-

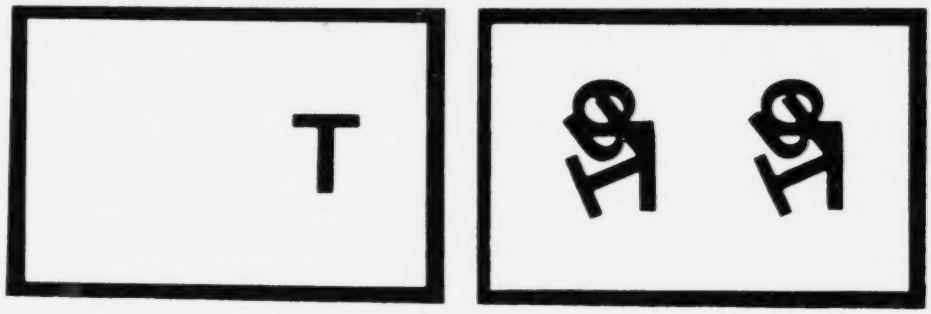


Fig. 1

Figure 1: The mask stimulus (right) and an example of the target stimuli. (The internal border represents the edge of the viewing field.)

cation under conditions of dichoptic presentation, e.g., if a target was presented to the temporal hemiretina of the left eye and the mask to the right eye. Both the target and mask durations were set at 10 msec to preclude eye movements. Luminances of both stimuli were set initially at 10 ft L, and a 50 percent Kodak neutral density filter was then used to reduce the mask luminance. Thus, the energy of the target stimuli was twice that of the mask.

A completely within Ss design was used, with each S receiving 10 targets to each hemiretina at each of 10 SOAs (0, 10, 20, 40, 50, 60, 80, 100, 150, and 200 msec). For half of the Ss, SOAs increased across trials and for the others, they decreased across trials. The 40 trials at each SOA were randomly divided among the four hemiretinae with the restriction that each hemiretina receive 10 presentations. On a particular trial, the S knew neither to which eye nor to which side of a constantly illuminated fixation point the stimulus was to appear.

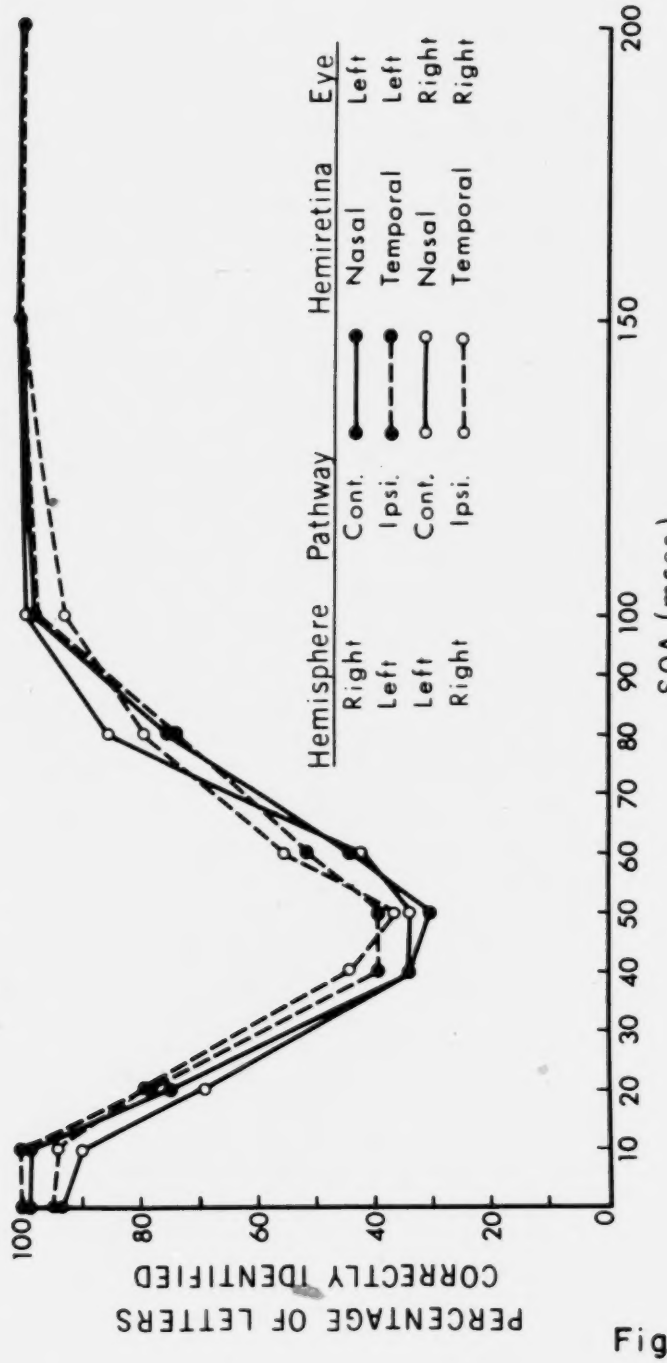


Fig. 2

Figure 2: Percentage of letters correctly identified as a function of SOA and hemiretina.

Eight Yale University undergraduates, naive about tachistoscopic viewing, were paid to serve as Ss. They were instructed to identify the letter and to guess if unsure. The entire procedure took about 40 minutes including a 5 minute rest at the halfway point.

Results and discussion. The number of letters correctly identified at each SOA was averaged across Ss for each hemiretina. These results are presented in Figure 2. An Eye X Pathway-type (Contralateral vs. Ipsilateral) X SOA X Ss analysis of variance revealed that only SOA was significant, $F(9,63)=16.6$, $p<.001$. All Ss demonstrated maximal masking at SOAs of either 40 or 50 msec. Examination of Figure 2 clearly shows U-shaped functions for each hemiretina but obviously neither the minima of these functions nor their ascending and descending components differed.

According to the hypotheses advanced earlier, two types of masking occurred. In peripheral processing, energy was the critical variable and at brief SOAs the higher energy target masked the after-coming mask; that is, the target won out in the competition for peripheral networks. At longer SOAs, the mask escaped peripheral impairment by the target and the rules of central processing took effect; namely, the mask had the advantage of being a second event and, as such, could disrupt the central processing of the target. On this account, the present results indicate that central masking did not differ as a function of the hemiretina to which the target and mask were delivered.

Another variable needs examination in the present context. Degree of eccentricity from the fixation point has been found to be a determinant of vowel reaction time (McKeever and Gill, 1972) and degree of metacontrast (Stewart and Purcell, 1970). We might suspect that degree of eccentricity affects peripheral and/or central masking. This possibility awaits investigation.

Finally, the existence of U-shaped functions in the present experiment reinforces the notion that they are not unique to the metacontrast situation as some have supposed (cf. Bridgeman, 1971).

REFERENCES

Bridgeman, B. (1971) Metacontrast and lateral inhibition. Psychol. Rev. 78, 528-539.

McKeever, W. F. and K. M. Gill. (1972) Interhemispheric transfer time for visual stimulus information varies as a function of retinal locus of stimulation. Psychon. Sci. 26, 308-310.

Purcell, D. G. and A. L. Stewart. (1970) U-shaped backward masking functions with nonmetacontrast paradigms. Psychon. Sci. 21, 361-363.

Stewart, A. L. and D. G. Purcell. (1970) U-shaped masking functions in visual backward masking: Effects of target configuration and retinal position. Percep. and Psychophys. 7, 253-256.

Turvey, M. T. (in press) On peripheral and central processes in vision: Inferences from an information processing analysis of masking with patterned stimuli. Psychol. Rev.

Weisstein, N. A. (1971) W-shaped and U-shaped functions obtained for monoptic and dichoptic disk-disk masking. Percep. and Psychophys. 9, 275-278.

Visual Storage or Visual Masking?: An Analysis of the "Retroactive Contour Enhancement" Effect

M. T. Turvey,[+] Claire Farley Michaels,[+] and Diane Kewley Port
Haskins Laboratories, New Haven

ABSTRACT

Standing and Dodwell (1972) reported that a contoured target stimulus, which is only poorly identified when exposed briefly against a steady background field, can be identified accurately if the field is terminated shortly after target offset. This observation was replicated and, in addition, it was shown that target identification is enhanced when the target onset is temporally proximate to the onset of the field. Furthermore, it was demonstrated that a continuous background field is not essential for either effect. It was argued that these "retroactive" and "proactive" enhancements of target identification were due to a complex interaction among forward, backward, and simultaneous masking.

INTRODUCTION

A target letter which is at or below recognition threshold when exposed briefly on a steady homogeneous or heterogeneous background field can become fully visible if the field terminates within about 100 msec of the target. This recent discovery of Standing and Dodwell (1972), which they have named retroactive contour enhancement (RCE), suggested to them a visual storage process for subliminal stimuli localized at a very early stage in the flow of the visual information.

The present paper examines the question of whether Standing and Dodwell's RCE phenomenon was the result of a storage process, as they have argued, or the result of some other kind of operation in the visual system.

EXPERIMENT I

The first experiment sought to replicate the basic finding of the Standing and Dodwell paper.

[+]Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

Method

Subjects. The subjects were four Yale University undergraduates who were paid $2.00 per hour for their services. All four subjects had normal or corrected-to-normal vision and all four were unfamiliar with tachistoscopic viewing.

Apparatus and stimulus materials. The same apparatus and the same materials were used for all four experiments reported. The stimuli were presented by means of a three-channel tachistoscope (Scientific Prototype, Model GB) with automatic slide changers. The viewing distance was 15 in. and the field of the tachistoscope subtended 3.5 deg vertical and 6.5 deg horizontal.

The target stimuli were a set of 100 trigrams constructed from the set of consonants with the restriction that no consonant was repeated within a trigram. The black letters on a white surround subtended .67 deg vertical and, on the average, .36 deg horizontal. The thickness of the letter parts subtended .13 deg visual angle, and the average separation between adjacent letters was .40 deg. The background field, or mask, was a random noise field, 3.5 deg vertical by 6.5 deg horizontal, used in previous experiments (see Turvey, in press, Figure 2). The random noise luminance was set at 0.6 ft L as measured by a SEI photometer.

Procedure. There were two durations of the random noise mask, 700 and 1000 msec (both had been used in the Standing and Dodwell experiments). Five intervals, 500, 600, 625, 650, and 675 msec, were used between onset of the mask and onset of the target stimulus. The duration of the target stimulus was 20 msec. Therefore, at the longest onset-onset interval of 675 msec the 700-msec mask terminated 5 msec after target offset and the 1000-msec mask terminated 305 msec after target offset. The mask exposure was superimposed on a fixation field of 0.02 ft L. The relation between the stimuli in the two conditions is shown in Figure 1.

Prior to testing each subject, the appropriate level of target stimulus luminance was determined so that the subject could identify an average 1.5 consonants in a trigram display exposed for 20 msec against a steady random noise background. The luminance value so determined was the target stimulus luminance used for the experiment. The average target luminance was 3.2 ft L.

Each subject was given twelve blocks of 20 trials, six with the random noise exposed for 700 msec and six with the random noise exposed for 1000 msec. Within a block the onset-onset times, or stimulus onset asynchronies (SOAs), were randomized, with the restriction that each SOA occurred four times. The subjects alternated between the random noise exposures of 700 and 1000 msec across the twelve blocks with two subjects beginning with the shorter duration and two with the longer. The consonant trigram changed with each trial, and the subjects were scored for the number of consonants correctly identified. All stimuli were presented monocularly to the right eye; Standing and Dodwell had used binocular presentation.

Results and Discussion

The function relating the proportion of consonants correctly reported to SOA for both exposure durations of the random noise are given in Figure 2.

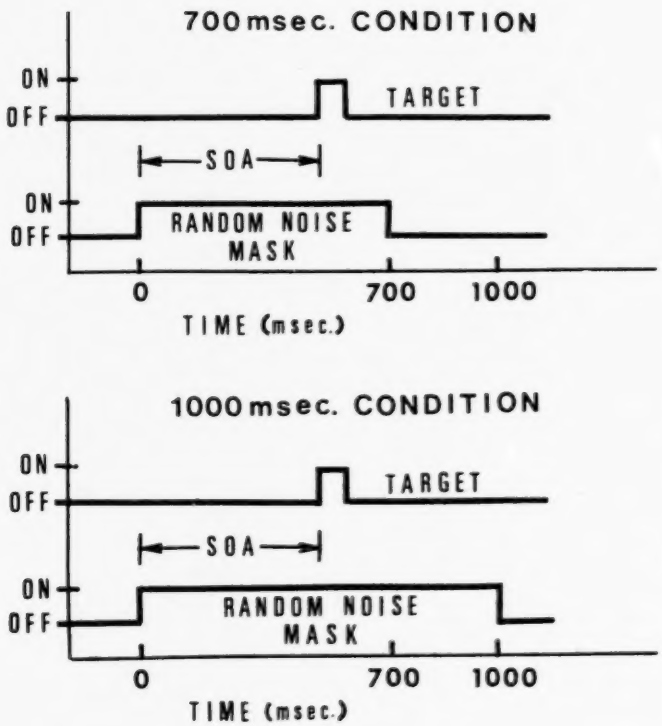


Fig. 1

Figure 1: Temporal relation between random noise mask and target stimuli in the 700 msec and 1000 msec conditions of Experiment I.

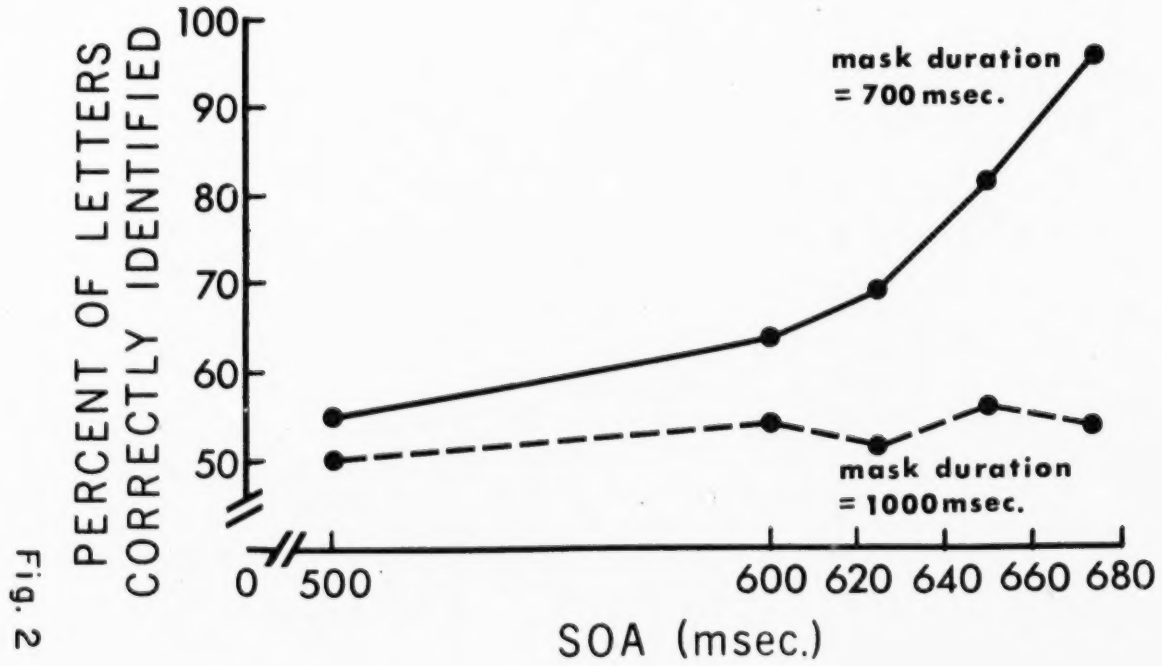


Fig. 2

Figure 2: Relation between percent correct identification and SOA with mask duration as the curve parameter in Experiment I.

A Treatment (random noise duration) x Treatment (SOA) x Subjects analysis yielded a significant effect of noise duration ($F = 74.87$; df = 1,3; $p < .005$), a significant effect of SOA ($F = 23.17$; df = 4,12; $p < .001$) and a significant interaction between random noise duration and SOA ($F = 9.57$; df = 4,12; $p < .005$).

The present experiment corroborates the main finding of the Standing and Dodwell experiments: the identification of a target stimulus increases when the background field on which the target stimulus is exposed terminates shortly after target-stimulus offset. In Standing and Dodwell's experiments the target stimuli consisted of one letter (S or L) in a forced-choice task, while the present experiment required identification of consonants presented in trigram strings. The implication is that the phenomenon is quite robust.

The only difference between the data of Standing and Dodwell and those presented here is that the increase in identification with increasing SOA in the 700-msec condition was more gradual in the present experiment. The source of this difference probably lies in the difference between the stimuli and response measures used in the two experiments.

EXPERIMENT II

The second experiment sought to determine whether the result of Experiment I could be obtained within a smaller range of mask exposures.

Method

The experiment was conducted in two parts and the apparatus and stimuli for both parts were the same as used in Experiment I. Both parts of the experiment used random noise exposures of 100 and 200 msec and a target exposure of 5 msec. The luminance of the random noise was set at 0.32 ft L for both mask durations and for both parts of the experiment. The stimuli were presented monocularly to the right eye.

Prior to testing of subjects in both Parts 1 and 2, the target luminance was determined at which approximately one item could be identified against a steady mask background. The target stimuli were then presented to each subject at this luminance for the course of the experiment. The average target luminance for the four subjects was 2.5 ft L.

Part 1. For each of the two mask exposures of 100 and 200 msec duration the target stimuli were superimposed on the mask field at SOAs of 10, 25, 50, 75, and 90 msec. The presentation of the mask exposure duration and SOA combinations followed the same pattern as described in Experiment I. However, only six blocks of 20 trials were used in the present situation, three for each mask duration. As before, SOAs were randomized within a block with each SOA occurring four times. Two of the authors were the subjects for this part of the experiment.

Part 2. The second part of the experiment differed from the first in that five additional SOAs of 110, 130, 150, 175, and 190 msec were examined at the mask exposure of 200 msec. The two subjects for this part of the experiment were Yale University undergraduates who had never participated in a tachistoscopic experiment before and who were paid for their services. Both subjects

received nine blocks of 20 trials; three blocks with the mask at 100 msec, three with the mask at 200 msec and SOAs less than 100 msec, and three blocks with the mask at 200 msec and SOAs greater than 100 msec. The order of the three blocks was partly counterbalanced across the two subjects and within a block SOAs were randomized with the restriction that each SOA was examined four times.

Results and Discussion

The averaged data of the two subjects in Part 1 are given in Figure 3 and those of Part 2 are given in Figure 4. Inspection of both figures reveals that the functions relating identification performance to SOA are nonmonotonic for the 100 msec mask exposure of Parts 1 and 2 and for the 200 msec exposure of Part 2. Thus, superimposing the target onto the mask background in close temporal proximity to either mask onset or mask offset led to enhancement in letter identification. In short, there was both proactive and retroactive facilitation and this U-shaped relation between target perceptibility and temporal location in the mask brings into question the memory interpretation of RCE proposed by Standing and Dodwell. According to that interpretation traces of the target and mask persist beyond their exposures, and, presumably, these perceptual traces decay exponentially with time. It is assumed that the target trace is the more durable of the two, either because the target is of greater intensity or because the target, unlike the mask, is contoured. In any event, if the mask offsets soon after the target exposure the mask trace will terminate before the target trace; consequently, the target trace may be accessed and the contour information recovered. On the other hand, if the mask offsets well beyond the target exposure, the mask or its perceptual trace may persist beyond the useful life of the target trace. Under these conditions recovery of the target would be impossible.

There are two fundamental difficulties with the persisting trace or visual storage hypothesis. First, by necessity it must predict a positively monotonic relation between target perception and SOA--in contrast to the U-shaped relation obtained in the present experiment. Second, although the storage hypothesis addresses itself to the improvement in target identification with proximity to mask offset, it does not speak, obviously, to the problem of why the steadily-presented mask should impede the perception of the target in the first place.

Two other points need to be made. First, in both Experiments I and II it was determined that the relation between the target and mask energies which produced RCE was such that the mask, at its longest durations of 1000 and 200 msec, did not significantly affect the accuracy of target identification if it followed the offset of the target or if it preceded the onset of the target. Second, while letter identification by three of the four subjects in the present experiment was poorest in the 100 msec mask duration condition at SOA = 50 msec, one subject's letter identification (in Part 2) was lowest at SOA = 30 msec and reasonably good at SOA = 50 msec. This subject variability accounts in part for the different minima observable in the 100 msec mask functions of Figures 3 and 4.

EXPERIMENT III

The third experiment was similar to Experiment II, with one major difference. Instead of overlapping the target and mask fields temporally, the mask was

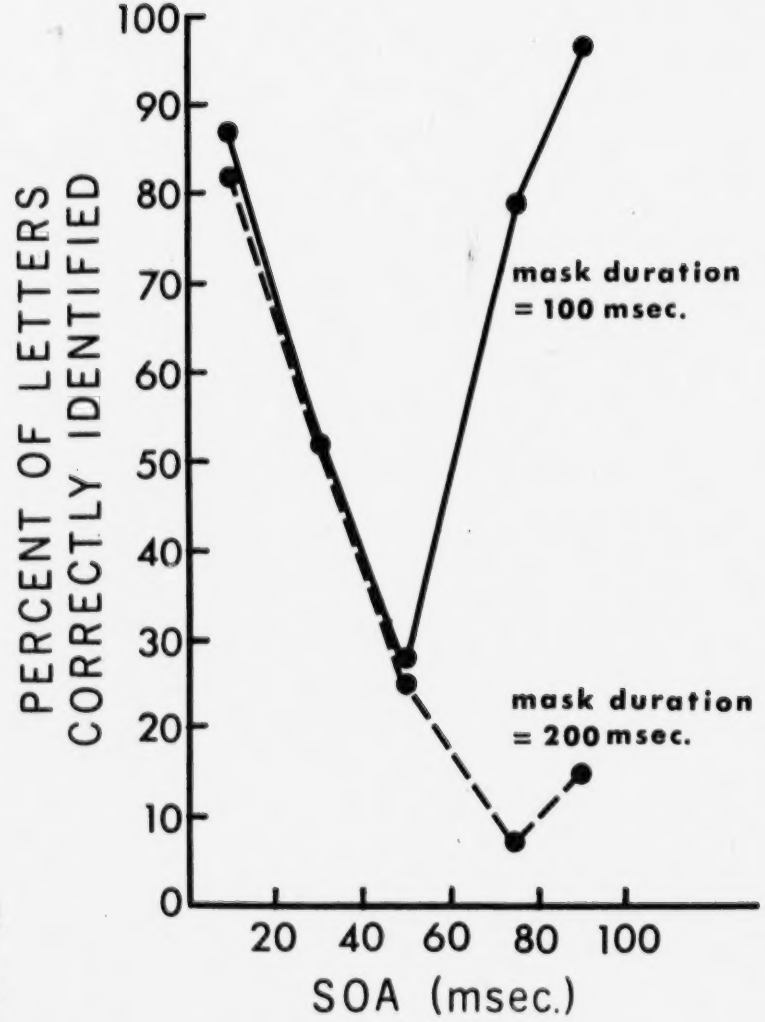


Fig. 3

Figure 3: Relation between percent correct identification and SOA with mask duration as the curve parameter in Part 1, Experiment II.

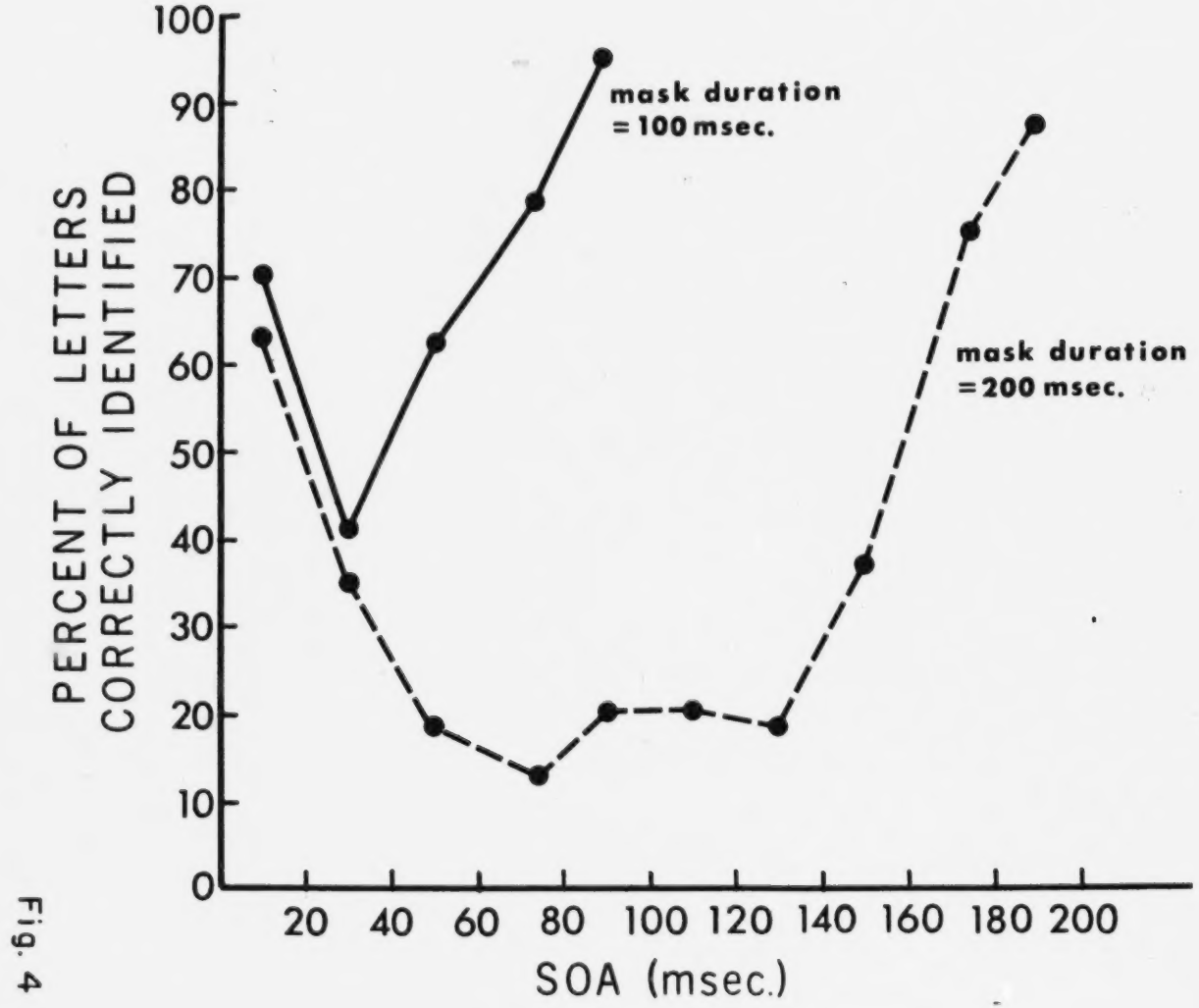


Figure 4: Relation between percent correct identification and SOA with mask duration as the curve parameter in Part 2, Experiment II.

terminated at target onset and continued from target offset. In other words, the target was inserted into a temporal "hole" in the mask, with the duration of this hole equal to the duration of the target. This was done to assess whether an uninterrupted presentation of the mask was essential to the proactive and retroactive facilitation effects demonstrated in Experiment II. A positive demonstration would argue that the paradigm used by Standing and Dodwell does not differ appreciably from the more common masking paradigm used to investigate interference between temporally discrete events.

Method

Two paid, tachistoscopically naive, Yale University undergraduates attempted to identify the trigram stimuli under conditions very much like those of Experiment II, Part 2. The general procedure was to present a target for 10 msec preceded and followed by random noise. Two identical random noise fields were presented on two separate channels of the tachistoscope. In this experiment, therefore, there was no fixation field since all three channels of the tachistoscope were used for the random noise/target/random noise sequence. The total duration of random noise/target/random noise was either 100 or 200 msec. In the 100 msec condition the target stimulus was presented at the following SOAs: 10, 30, 50, 75, and 90 msec, where the SOA was measured from the onset of the first exposure of the random noise. In the 200 msec condition the target was presented at five additional SOAs of 110, 130, 150, 175, and 190 msec. Since the random noise mask was off for the duration of the target, the total time the mask was exposed was 90 msec in the 100 msec condition and 190 msec in the 200 msec condition.

Both subjects received nine blocks of 20 trials in the fashion described in Part 2 of Experiment II. Prior to the experiment the target luminance was determined by an identification accuracy of a little less than one consonant per trigram exposure when the target was superimposed upon the continuous mask. This luminance was 4.0 ft L for both subjects. The luminance of both random noise fields was 0.32 ft L and the stimuli were presented to the right eye.

Results and Discussion

The averaged data for the two subjects are plotted in Figure 5. Comparison of Figure 5 with Figure 4 shows that the functions relating SOA to consonant identification for the 100 and 200 msec conditions of the present experiment are identical in form to those of Experiment II. The conclusion we draw from this is that a continuous mask is essential neither for the RCE effect nor for the proactive facilitation demonstrated in Experiment II.

For the luminance levels used in the present experiment it was determined that with the mask only preceding or only following the target, impairment in the identification of the target was minimal. This was true for both the shortest (10 msec) and the longest (190 msec) exposures examined. Uttal (1969) has recently reported an experiment which showed that a leading mask and a lagging mask which failed at a certain interval to impair target identification when presented separately, significantly reduced target identification when presented in combination. Uttal (1969, 1971) and Walsh (1971) have speculated that the elevated masking evident in Uttal's (1969) combined forward and backward masking situation may have resulted from the summation of "latent" masking effects which in themselves were inadequate to affect target recognition.

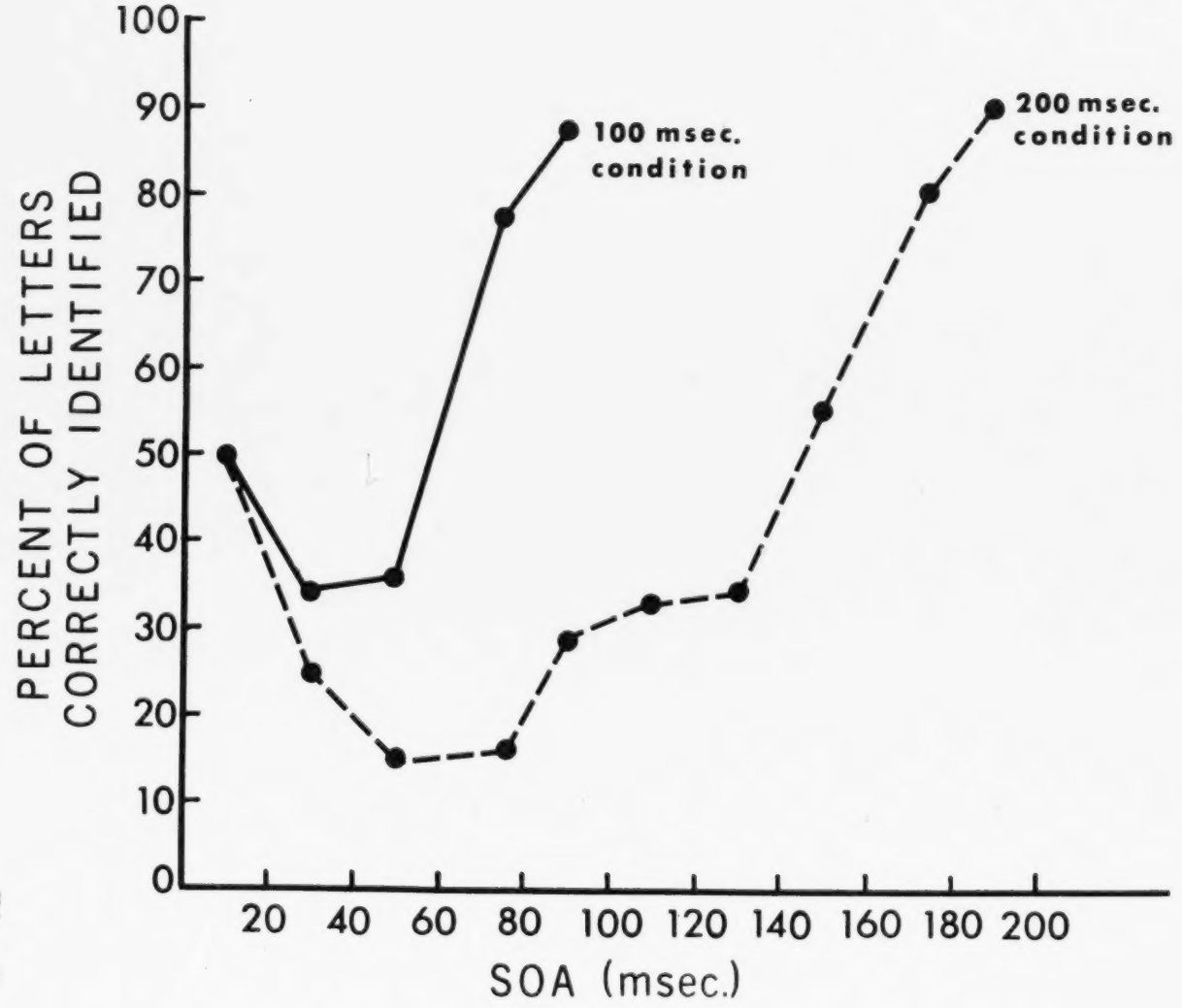


Fig. 5

Figure 5: Relation between percent correct identification and SOA with total mask-target-mask time as the curve parameter in Experiment III.

If target perception in the present experiment was determined by the joint effects of "latent" forward and backward masking, then we can argue that the systematic, nonmonotonic changes in target perception as a function of SOA were the result of systematic changes in the differential masking influences of the forward and backward exposures of the random noise. We should recall at this point that in the present experiment SOA was defined as the interval elapsing between the onset of the first exposure of the random noise and the onset of the target. Therefore, since mask/target/mask time was held constant within a condition, increasing SOA was equivalent to increasing the duration of the forward mask and to decreasing the duration of the backward mask.

Within limits, an essential determinant of monoptic masking is the energy relation between the stimuli, and in the view of a recent discussion of masking (Turvey, in press), when two stimuli are competing for common peripheral networks, the stimulus of greater energy tends to have the advantage. In the present experiment the target was more intense than the mask, 4.0 ft L compared to 0.32 ft L, and therefore, for equivalent exposure durations the target was of greater energy than the mask, where energy is defined as: duration x luminance. Thus, at brief SOAs in the present experiment the target probably impaired the leading random noise more than the leading random noise impaired the target. Therefore, we suspect that at these brief SOAs forward masking effects were minimal, and the weight of the masking action was on the lagging exposure of the random noise. This, we may recall, was not an effective masker when presented in the absence of the leading exposure. However, at longer SOAs, e.g., 60-70 msec, the forward random noise, because of its increased energy, was a more pronounced forward masker and could more effectively interact with the lagging random noise to impede target perception. In short, the transition from brief to moderately long SOAs in the present experiment was accompanied by an increase in the effectiveness of the forward masker. In the presence of the backward mask this resulted in the decreasing perceptibility of the target as a function of SOA. This accounts for the "proactive facilitation" effect evident in the present data, and we can account for the RCE effect in a similar fashion.

We can assume that at the longer SOAs the duration, and therefore, the energy of the backward mask was reduced below that point at which independently it could maximally influence the target. We should bear in mind, of course, that the maximal influence of the random noise as either an independent forward or backward masker was not sufficient to impair target identification. As Walsh (1971:265) has commented, "discriminability is not equivalent to invulnerability." In any event, at the longest SOAs the backward masking action of the random noise was minimized, leaving the leading exposure of the random noise as the major source of masking. Consequently, in both conditions of the present experiment target identification increased as backward masking decreased, with increases in SOA from the middle to the longest values. This increase in target identification at the longest SOAs is RCE.

Thus, the RCE effect manifest in the present experiment may be interpreted as a change in target identification resulting from variation in the joint masking effect of leading and lagging masks. We believe that this conclusion can be generalized to the RCE effect observed in Experiments I and II of the present paper, and to the experiments of Standing and Dodwell. In those experiments the target and mask overlapped temporally, a condition which, in view of the foregoing, can be interpreted as a target/mask composite preceded and

followed by a mask. At most, performance in a continuous mask situation should be poorer than performance in a comparable, interrupted mask situation in view of the additional element of latent simultaneous masking. In our view, therefore, the difference between the two situations is simply one of degree.

EXPERIMENT IV

The fourth experiment was designed to test the masking interpretation of Experiment III and thus, of the RCE effect. If target identification in Experiment III was due to the joint effect of the forward and backward mask, then varying the masking capability of either the leading or lagging exposure of random noise, or both, should affect target identification.

We have good reason to believe that in the present series of experiments the locus of the influence of the random noise mask on target stimulus perception was peripheral rather than central. It was determined that at the intensity levels used in Experiments I, II, and III, the random noise, either overlapping or not overlapping temporally with the target, could not impede target perception when it was presented to one eye and the target stimulus was presented to the other. In addition, the random noise mask of the present series of experiments had been shown previously, over a range of conditions, to be a relatively ineffective dichoptic masker for the present set of trigram target stimuli (Turvey, in press).

Accepting the peripheral origin of the RCE effect demonstrated in the three experiments of the present paper and, we presume, in the experiments of Standing and Dodwell, we would expect that a major determinant of peripheral masking, the energy relation between the stimuli (Turvey, in press) is also a major determinant of RCE. In this view, what was important to the RCE effect of Experiment III was the relation between the energies of the pre- and post-target mask exposures and the target, and not the proximity of the target to mask offset. To test this hypothesis the luminance and duration of the lagging random noise exposure were varied. If energy, rather than the time before mask offset, was the important variable determining RCE, then target perceptibility should not be altered by the reciprocal interchange of mask duration and mask intensity, but it should be significantly influenced by the independent manipulation of either.

Method

Three Yale University undergraduates, naive about tachistoscopic presentation, were paid for their participation. The experiment used the target and mask stimuli of the preceding experiments.

In most respects the experimental procedure was similar to that described for Experiment III, i.e., a 10 msec target was preceded and followed by a random mask exposure. In the present experiment, however, the duration and intensity of the lagging exposure were systematically varied. The luminance of the lagging random noise was set at 3.2 ft L and then presented at 50% (1.6 ft L), 25% (0.8 ft L), or 10% (0.32 ft L) of this value. These three luminance levels of the lagging mask were combined factorially with three durations of 10, 20, and 50 msec. Each subject received six blocks of 15 trials, two blocks per mask luminance level. Within each block five trigrams were followed by a 10 msec mask exposure, five by a 20 msec mask exposure, and five

by a 50 msec mask exposure, with the mask duration randomized within the 15 trials. The six blocks were balanced across the three subjects such that each luminance level appeared with equal frequency at each position in the test. Thus, each subject received all nine luminance/duration conditions with luminance level counterbalanced and duration randomized within luminance level.

Throughout the experiment the leading random noise exposure was constant at 50 msec duration and 0.32 ft L intensity. The mask was terminated at target onset and continued from target offset. The same target luminance of 3.2 ft L was used for the three subjects, and all stimuli were presented monoptically, to the right eye.

Results and Discussion

The proportion of letters correctly reported at each luminance/duration combination is given in Table 1. Inspection of the table shows that increasing either the duration or the intensity of the lagging mask decreased target perceptibility. An analysis of variance showed that both main effects of duration

TABLE 1

EXPERIMENT IV: Proportion of letters correctly identified as a function of luminance and duration of lagging mask.

Duration (msec)	Luminance (per cent of 3.2 ft L)		
	10%	25%	50%
10	.86	.83	.42
20	.77	.51	.11
50	.41	.17	.07

and of luminance were significant; $F = 42.13$; df = 2,4; $p < .01$ and $F = 129.20$; df = 2,4; $p < .01$, respectively. Moreover, it was obvious that the total energy (luminance x duration) of the lagging mask was the important determinant of performance in the present experiment. First, increasing both duration and intensity resulted in a greater impairment in target identification than increasing either independently. Second, in the three cells of Table 1 in which energy was held constant by the reciprocal variation of luminance and duration, i.e., 10% x 50 msec, 25% x 20 msec, and 50% x 10 msec, target perceptibility was relatively constant. We should also note that reducing mask duration at each of the three luminance levels resulted in an improvement in target identification; thus, the RCE effect was observed at each luminance level.

On the evidence of Experiment IV we may conclude that the enhancement in target identification that accompanied the reduction in the target offset/mask offset interval of Experiments I-III was due to the reduced energy of the mask exposure following the target rather than to the increased proximity of offsets.

But more generally, we may conclude that the nonmonotonic relation between target perception and SOA evidenced in the present experiments was due to systematic changes in the joint masking effect of the preceding and succeeding exposure of the random noise.

REFERENCES

Standing, L. G. and P. C. Dodwell. (1972) Retroactive contour enhancement: A new visual storage effect. Quart. J. Exp. Psychol. 24, 21-29.

Turvey, M. T. (in press) On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. Psychol. Rev.

Uttal, W. R. (1969) The character in the hole experiment: Interaction of forward and backward masking of alphabetic character recognition by dynamic visual noise. Percept. Psychophys. 6, 177-181.

Uttal, W. R. (1971) A reply to Walsh. Percept. Psychophys. 10, 267-268.

Walsh, T. (1971) The joint effect of forward and backward visual masking: Some comments on Uttal's "Character in the hole" experiment. Percept. Psychophys. 10, 265-266.

Reading and the Awareness of Linguistic Segments

Isabelle Y. Liberman,[+] Donald Shankweiler,[+] Bonnie Carter,[++] and F. William Fischer[++]

ABSTRACT

Since many children who can understand spoken language cannot learn to read, we have asked what the child needs for reading beyond that which he already commands for speech. One important extra requirement is conscious awareness of phonemic segmentation. In speech, phonemic segments are normally encoded into units of approximately syllabic size. Awareness of linguistic structure at the phoneme level might therefore be difficult to attain, more difficult in any case than at the syllable level. Using a task which required our four-, five-, and six-year-old subjects to tap out the number of segments in spoken items, we found that analysis into phonemes is, indeed, significantly harder than analysis into syllables. At all three ages, far fewer children reached criterion with the phoneme task; those who achieved criterion required a greater number of trials to do so.

There are many children who readily acquire the capacity to speak and understand language but do not learn to read and write it. It is of interest, therefore, to ask what is required in reading a language that is not required in speaking or listening to it. The first answer which comes to mind, of course, is that reading requires visual identification of optical shapes. Since our concern here is with reading an alphabetic script, we may well ask whether the rapid identification of letters poses a major obstacle for children learning to read. The answer is that for most children perception of letter shapes does not appear to be a serious problem. There is considerable agreement among investigators that by the end of the first year of school, even those children who make little further progress in learning to read generally show no significant difficulty in the visual identification of letters as such (Vernon, 1960; Shankweiler, 1964; Doehring, 1968; I. Liberman, Shankweiler, Orlando, Harris, and Berti, 1971; Kolers, 1972).

[+]University of Connecticut, Storrs, and Haskins Laboratories, New Haven.

[++]University of Connecticut, Storrs.

Acknowledgment: The authors are indebted to Dr. A. M. Liberman, Haskins Laboratories, for many helpful suggestions and a critical reading of the text.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

Beyond identification of letters, learning to read requires mastery of a system which maps the letters to units of speech. There is no evidence, however, that children have special difficulty in grasping the principle that letters stand for sounds. Indeed, children can generally make appropriate sounds in response to single letters, but are unable to proceed when they encounter the same letters in the context of words (Vernon, 1960).

Another possible source of difficulty is that the relation in English between spelling and the language is often complex and sometimes highly irregular. But even when the items to be read are carefully chosen so as to include only those words which map the sound in a simple, consistent way and which are part of their active vocabularies, many children continue to have difficulties (Savin, 1972).

There remains at least one other possible barrier to reading acquisition. As suggested by several investigators (I. Liberman, 1971; Mattingly, 1972; Savin, 1972; Shankweiler and I. Liberman, 1972), in order to read an alphabetically written language, though not necessarily to speak and listen to it, the child must be quite consciously aware of phonemic segmentation. Let us consider the child trying to read the orthographically regular word bag. We will assume that he can see the written word and can identify the letters b, a, and g. We will assume also that he knows the sounds of the individual letters, which he might say as [bʌ] [ae] [gʌ]. But if that is all he knows, then he would presumably read the word as the trisyllable "buhaguh," which is a nonsense word and not the meaningful monosyllable "bag." If the child is to map the three letters of the printed word bag onto the one-syllable spoken word "bag" that he already knows, he must be consciously aware that the spoken word consists of three phonemic segments.

As we have said earlier, we believe that it is this requirement, this need to be consciously aware of the phonemic segmentation of the spoken word, that presents real difficulties for many children learning to read. But why should this pose special difficulties? If the sounds of speech bore a simple one-to-one relation to the phonemic units in the spoken message, just as the letters do (at least in the orthographically regular case), it would be hard to see why the child should be unaware of the phonemic segmentation. That is, if there were in the word "bag" three acoustic segments, one for each of the three phonemes, then the segmentation of the word that is represented in its spelling would presumably be quite apparent.

However, as extensive research in speech perception has shown (Fant, 1962; A. Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Stevens, 1972), the segmentation of the acoustic signal does not correspond directly or in any easily determined way to the segmentation at the phonemic level. It should be emphasized that this lack of correspondence does not come about simply because the sounds of the phonemes are joined together, as are the letters of the alphabet in cursive writing or as may be implied by the reading teacher who urges the child to blend "buhaguh" into "bag." Rather, the phonemic segments are truly encoded in the sound. In the case of "bag," for example, the initial and final consonants are folded into the medial vowel, with the result that information about the successive phonemic segments is transmitted more or less simultaneously on the same parts of the sound. In exactly that sense, the syllable "bag" is not three acoustic segments, but one. This is not to say that the phonemic elements are not real, but only that the relation between them and the

sound is that of a very complex code, not a simple substitution cipher (A. Liberman et al., 1967). To recover the phonemic segments requires a correspondingly complex decoding process. In the normal course of perceiving speech, these processes go on quite automatically and, in the usual case, without conscious awareness.

That it might be more than a little difficult to bring the processes of phonemic analysis above the level of conscious awareness is suggested by the fact that an alphabetic method of writing has been invented only once (Gelb, 1963) and is a comparatively recent development in the history of writing systems. Of more immediate relevance to us is the evidence that children with reading disabilities may have difficulty even with spoken language when required to perform tasks that might demand explicit awareness of phonemic structure. These children are often reported, for example, to be deficient in rhyming, in recognizing that two different monosyllables share the same first (or last) phonemic segment (Monroe, 1932) and according to recent research (Savin, 1972), in speaking Pig Latin, which demands a conscious shift of the initial phonemic segment to the final position in the word.

As noted earlier, research on speech perception has found that the acoustic unit into which the phonemic elements in speech are encoded is of approximately syllabic dimensions. We would therefore suppose that the number of syllables (though not necessarily the location of syllable boundaries) might be more readily available to consciousness than the phonemes. If so, we might then have an explanation for the assertion (Makita, 1968) that Japanese kana, which is approximately a syllabary, is easier for the child to master. Since word segments are perhaps even more accessible, we might expect that an orthography which represents each word with a different character, as in the case of Chinese or the closely related Japanese kanji, would also not cause, in the beginning reader,[1] the particular difficulties that arise in mastering the more analytic alphabetic system. Indirect evidence of the special burden imposed on the beginning reader by an alphabetic script can be found also in the relative case with which reading-disabled children learn kanji-like representations of language while being unable to break the alphabetic cipher (Rozin, Poritsky, and Solsky, 1971). It is worth noting, in the context of the foregoing observations, that since the time of the Greeks, methods of reading instruction (Mathews, 1966), have sporadically reflected the assumption on the part of educators that the phonemic structure is more easily taught through the use of syllabic units, presumably because the latter are easier for the child to apprehend.

---

[1] It should be emphasized that the advantage of a logographic script is limited to the beginning reader. For the older child and adult, the kanji system presents other difficulties, such as the large number of characters to be learned (some 1,800 kanji for the daily newspaper, 5,000 for a novel). As to the Japanese kana, it appears an ideal writing system for the open-syllable Japanese language with its relatively small number of syllables (approximately 90) but would be hardly appropriate for the complex and highly variable syllable structure of English. Though neither the logograph nor the syllable would be recommended as substitutes for the alphabet in the English writing system, they might be considered for use as units in initial teaching methods. L. Gleitman and P. Rozin of the University of Pennsylvania (personal communication) have incorporated both into a teaching method which they consider to be promising with problem readers.

However, no research has been addressed specifically to the question of whether children, when they begin to read, do, in fact, find it difficult to make an explicit phonemic analysis of the spoken word and whether this ability comes later and is more difficult than syllabic analysis. In this study, we will see how well children at nursery school, kindergarten, and first grade ages can identify the number of phonemic segments in spoken words and will compare this with their ability to deal similarly with syllables.

METHOD

Subjects

The subjects were 135 white, middle class children from a public preschool program in the suburban town of Manchester, Connecticut, and from the elementary school in the adjoining town of Andover, Connecticut. They included 46 nursery schoolers ages 48 to 68 months, mean age 58 months (S.D. 5.40), 49 kindergarteners aged 63 to 79 months, mean age 70 months (S.D. 4.10), and 40 first graders aged 65 to 96 months, mean age 83 months (S.D. 5.50). The nursery school group contained 21 boys and 25 girls; the kindergarteners, 18 boys and 31 girls; the first graders, 15 boys and 25 girls. All available children at the appropriate grade levels in the participating schools were used, with the following exceptions: among the nursery school children, four with speech and hearing problems, 12 who refused to enter into the testing situation at all, and five who were so inattentive and distractible that demonstration trials could not be carried out; among the kindergarteners, one who had returned to kindergarten after several months in first grade and one whose protocol was spoiled by equivocal responses. No first graders were excluded.

Alphabetized class registers at each grade level were used to alternate the children between the two experimental groups, the one requiring phoneme segmentation (Group P) and the other, syllable segmentation (Group S). Equalization of the numbers of children assigned to each type of task was complicated at the nursery school level by the sporadic lack of participation by individual children. An attempt to equalize the numbers of boys and girls in the two task groups was hampered by the unequal numbers of the two sexes at all grade levels. The final composition of the groups is shown in Table 1.

TABLE 1: Composition of phoneme (P) and syllable (S) groups across grade and sex.

Grade	Nursery School		Kindergarten		First Grade	
Task	P	S	P	S	P	S
Male	9	12	9	9	7	8
Female	11	14	15	16	13	12
Total	20	26	24	25	20	20

The level of intelligence of all the subjects was assessed by the Goodenough Draw-A-Person Test (DAP). When computed across tasks, the mean DAP IQ was 110.06 (S.D. 18.20) for the syllable group and 109.19 (S.D. 15.73) for the phoneme group. Across grade levels, the mean IQ was 112.11 (S.D. 17.04) for the nursery schoolers, 108.90 (S.D. 17.92) for the kindergarteners, and 107.73 (S.D. 15.90) for the first graders. Two-way analyses of variance performed on the DAP IQ scores revealed no significant differences in IQ, either across tasks or across grade levels. In addition, the mean chronological ages of the two task groups were also found to be not significantly different. The mean age in months of the syllable group was 69.41 (S.D. 11.25); of the phoneme group, 69.58 (S.D. 11.18). Therefore, any performance differences in the two types of segmentation can reasonably be taken to be due to differences in the difficulty of the two tasks.

Procedure

Under the guise of a "tapping game," the child was required to repeat a word or sound spoken by the examiner and to indicate, by tapping a small wooden dowel on the table, the number (from one to three) of segments (phonemes in Group P and syllables in Group S) in the stimulus items. Four sets of training trials containing three items each were given. During training, each set of three items was first demonstrated in an order of increasing complexity (from one to three segments). When the child was able to repeat and tap each item in the triad set correctly, as demonstrated in the initial order of presentation, the items of the triad were presented individually in scrambled order without prior demonstration and the child's tapping corrected as needed. The test trials, which followed the four sets of training trials, consisted of 42 randomly assorted individual items of one, two, or three segments which were presented without prior demonstration and corrected by the examiner, as needed, immediately after the child's response. Testing was continued through all 42 items or until the child reached criterion of tapping six consecutive items correctly without demonstration. Each child was tested individually by the same examiner in a single session during either late May or June, 1972.

Instructions given to the two experimental groups at all three grade levels were identical except that the training and test items involved phonemic segmentation in Group P and syllabic segmentation in Group S. (See Stimulus Materials for further details.) The instructions used for the syllable task were as follows:

> "We are going to play a tapping game today. I'm going to say some words and sounds and tap them after I say them. Listen, so you'll see how to play the game."
>
> Training trials.
>
> Step 1. (Examiner demonstrates with first training triad.) "But (one tap). Butter (two taps). Butterfly (three taps)."
>
> "Now, I want you to do it. Say but...Good. Now, tap it... Good. Now, put your stick down. Say butter...Tap it... Good. Now, put your stick down. Say butterfly...Tap it... Now, put your stick down." (If the child makes an error in tapping, the entire triad is demonstrated again. If error persists, E goes on to Step 3. If tapping is correct, E goes to Step 2.)

Step 2. "Now, let's do it again to make sure you've got the idea. I'll mix them up and see if I can catch you. Say butter... Now, tap it...Say but...Now, tap it...Say butterfly...Now, tap it."

Step 3. "Let's try some more words. I'll do it first." (Demonstration is continued with the next three training triads, following all procedures in Steps 1 and 2 as needed.)

Test trials.

"Now, we'll play the real game. I'll say a word, but I won't tap it, because you know how to play the game yourself. So, you say the word after me and then tap it. After each word, be sure to put your stick down so I'll know you've finished tapping."

"Here's the first word. _____. You say it and tap it." If the child taps incorrectly, E says, "Listen to the way I do it. Now, you do it the same way I did it." If the child still taps incorrectly, E says, "Okay, here's the next one," and goes on to the next word. If the child taps correctly, E says, "Good! Here's the next one."

The same procedure is continued until the end of the list of 42 items or until the child reaches criterion of tapping six consecutive items correctly without demonstration.

Stimulus Materials

The training trials for the phoneme task included the following four triads:

1) /u/ (as in moo)
   boo
   boot

2) /ae/ (as in hat)
   as
   has

3) /O/ (as in go)
   toe
   tall

4) /I/ (as in bit)
   ma
   cut

For the syllable task, the training trials were:

1) but
   butter
   butterfly

2) tell
   telling
   telephone

3) doll
   dolly
   lollipop

4) top
   water
   elephant

It will be noted that in both the Group P and Group S training trials, the first two triads were formed by adding a segment to the previous item, while in the third triad, the final item varied from this rule. In the fourth triad,

all three items varied in linguistic content, so as better to prepare the child for the random distribution of linguistic elements in the test trials.

TABLE 2: Test list for the phoneme segmentation task.

1.	is	15.	/ɔ/ (as in bought)	29.	/U/ (as in bull)
2.	/ɛ/ (as in bet)	16.	cough	30.	toys
3.	my	17.	pot	31.	cake
4.	toy	18.	/u/ (as in boot)	32.	cool
5.	/ae/ (as in bat)	19.	heat	33.	/e/ (as in bait)
6.	/i/ (as in beet)	20.	he	34.	Ed
7.	soap	21.	/a/ (as in hot)	35.	cup
8.	/I/ (as in bit)	22.	pa	36.	at
9.	his	23.	mat	37.	book
10.	pout	24.	/ʌ/ (as in but)	38.	/Uk/ (as in book)
11.	mine	25.	so	39.	lay
12.	caw	26.	/ai/ (as in bite)	40.	coo
13.	out	27.	up	41.	/O/ (as in boat)
14.	red	28.	/au/ (as in bout)	42.	oy

TABLE 3: Test list for the syllable segmentation task.

1.	popsicle	15.	chicken	29.	father
2.	dinner	16.	letter	30.	holiday
3.	penny	17.	jump	31.	yellow
4.	house	18.	morning	32.	cake
5.	valentine	19.	dog	33.	fix
6.	open	20.	monkey	34.	bread
7.	box	21.	anything	35.	overshoe
8.	cook	22.	wind	36.	pocketbook
9.	birthday	23.	nobody	37.	shoe
10.	president	24.	wagon	38.	pencil
11.	bicycle	25.	cucumber	39.	superman
12.	typewriter	26.	apple	40.	rude
13.	green	27.	funny	41.	grass
14.	gasoline	28.	boat	42.	fingernail

As can be seen in Tables 2 and 3, both experimental test lists contained an equal number of randomly distributed one-, two-, and three-segment items. These were presented in the same order to all children in each experimental group. The items had been checked against word recognition and vocabulary tests to insure that they were reasonably appropriate for the vocabulary level of the children. In addition, a pilot study carried out in a day-care center had confirmed the suitability of both the vocabulary level and the test procedure for children aged three to six years. No further control of linguistic content was attempted in the Group S items, except that the accent in the two- and three-segment items was always on the first syllable. In the Group P

list, an effort was made to include as many real words, rather than nonsense words, as possible. Of necessity, the one-segment items, which consisted of 14 different vowel sounds, usually formed nonwords. The two-segment items in Group P were constructed by adding a consonant in the initial position to six of the vowels and in the final position to the remaining eight vowels. All of the three-segment items in Group P, with one exception, were constructed by the addition of one consonant to a two-segment item in the list.

RESULTS

The number of trials taken by each child to reach criterion level (six correct test trials without demonstration by the examiner) is displayed in Table 4 for the phoneme (P) and syllable (S) task groups at three grade levels.

TABLE 4: Number of trials taken by each child to reach criterion in the phoneme (P) and syllable (S) groups at three grade levels. (Maximum possible trials is 42. Blanks represent children who did not reach criterion.)

Grade	Nursery School (age four)		Kindergarten (age five)		First Grade (age six)	
Task	P	S	P	S	P	S
	—	—	—	—	—	—
	—	—	—	—	—	—
	—	—	—	—	—	27
	—	—	—	—	—	19
	—	—	—	—	—	18
	—	—	—	—	—	13
	—	—	—	—	41	10
	—	—	—	—	40	10
	—	—	—	—	36	10
	—	—	—	—	35	9
	—	—	—	—	34	6
	—	—	—	—	34	6
	—	—	—	—	28	6
	—	—	—	27	28	6
	—	42	—	25	22	6
	—	31	—	19	19	6
	—	36	—	13	13	6
	—	36	—	13	10	6
	—	35	—	10	10	6
	—	29	—	7	9	6
		25	38	7		
		25	25	6		
		16	23	6		
		12	18	6		
		6		6		
		6				
Mean number trials to reach criterion	—	25.7	26.0	12.1	25.6	9.8

It is apparent that the test items were more readily segmented into syllables than into phonemes. In the first place, we see from the table that the number of children who were able to reach criterion was markedly greater in the syllable group than in the phoneme group, whatever the grade level. This aspect of performance is shown graphically in Figure 1 in terms of the percentages of children in nursery school (age four), kindergarten (age five), and first grade (age six) who reached criterion in the two types of segmental analysis. One can see that at age four, none of the children could segment by phonemes, while nearly half (46%) could segment by syllables. Ability to perform phoneme segmentation did not appear at all until age five and then it was demonstrated by only 17% of the children; in contrast, almost half (48%) of the children at that age could segment syllabically. Even at age six, only 70% succeeded in phoneme segmentation, while 90% were successful in the syllable task.

If we now refer back to Table 4 we see that the relatively greater difficulty of phoneme segmentation is indicated not only by the fact that fewer children reached criterion level with the phoneme task than with the syllables, but also by the fact that those children who did reach criterion on the phoneme task took a greater number of trials to do so. The mean number of trials to reach criterion in syllable segmentation was 25.7 at age four, 12.1 at age five, and 9.8 at age six. For the phoneme segmentation group, on the other hand, we cannot say what the mean number of trials was at age four, since no child reached criterion at that age. At age five, only four children succeeded in phoneme segmentation, and their mean number of trials was 26.0, more than twice the mean of the 12 children of the same age who completed the syllable task. At age six, 14 children met the criterion in phoneme segmentation, and here the mean of 25.6 is nearly three times that of the 18 children who succeeded on the syllable task. Moreover, the mean of the phoneme group at age six is roughly equal to the mean of the syllable group at age four.

The contrast in difficulty between the two tasks can also be seen in Table 4 in terms of the number of children who achieved criterion level in six trials, which, under the procedures of the experiment, was the minimum possible number. For the children who worked at the syllable task, the percentage who reached criterion in the minimum time increased steadily over the three age levels. It was 7% at age four, 16% at age five, and 50% at age six. In striking contrast to this, we find that in the phoneme group no child at any grade level attained the criterion in the minimum time.

An analysis of variance was carried out to assess the contribution of the several conditions of the experiment. The measure on which the analysis was performed was the mean number of trials taken to reach criterion. For all those children who did not reach criterion, we here assigned an arbitrary score of 43, which is one more than the 42-trial minimum provided by the procedures of the experiment. Due to the unequal numbers of subjects in each cell and the necessity of retaining all the data, the harmonic mean was used in the computation (Lindquist, 1940). The three variables considered were task, grade, and sex. The analysis of variance for these variables and their interactions is summarized in Table 5.

TABLE 5: Analysis of variance summary table--main effects and interactions of task, grade level, and sex.

Source	df	MS	F
Total	134		
Task (T)	1	5053.61	39.50**
Grade (G)	2	2718.91	21.25**
Sex (S)	1	693.18	5.42*
T x G	2	280.58	2.19
T x S	1	40.29	0.31
G x S	2	415.20	3.25*
T x G x S	2	19.24	0.15
Error	123	127.95	

*$p < .05$
**$p < .001$

It can be seen that the main effect of task was significant with $p < .001$. The same high level of significance ($p < .001$) was also found for the effect of grade. Somewhat less significant effects indicate that girls were superior to boys ($p < .05$) and that there was a grade-by-sex interaction ($p < .05$). Inspection of the test data suggested that the grade-by-sex interaction could be attributed mainly to the superior performance of first grade girls. T-tests showed no significant differences between boys and girls at the nursery school and kindergarten ages in either the phoneme or syllable tasks, but at the first grade level, the girls were superior in the syllable task ($p < .02$) and also, though less significantly, in dealing with the phonemes ($p < .10$).

DISCUSSION

We have suggested that one way in which reading an alphabetic script differs from perceiving speech is that reading, but not speech perception, requires an explicit awareness of phoneme segmentation. In our view, the awareness of this aspect of language structure might be particularly difficult to achieve because there is in the speech signal no simple and direct reflection of phonemic structure. Phonemic elements are encoded at the level of sound into units of syllabic size. It ought, therefore, to be easier to become aware of syllables.[2] In this study, we have found that analysis of a word into phonemes is indeed significantly more difficult than analysis into syllables at ages four, five, and six. Far fewer children in the groups which received the phoneme task were able to reach criterion level; those who did, required a greater number of trials; and none achieved criterion in the minimum time.

[2] P. MacNeilage (personal communication) has suggested that this is true of educated adults as well. In a recent experiment, he has found that his subjects show virtually perfect agreement as to the number of syllables words contain, but considerable variability in their judgments of the number of phonemic constituents.

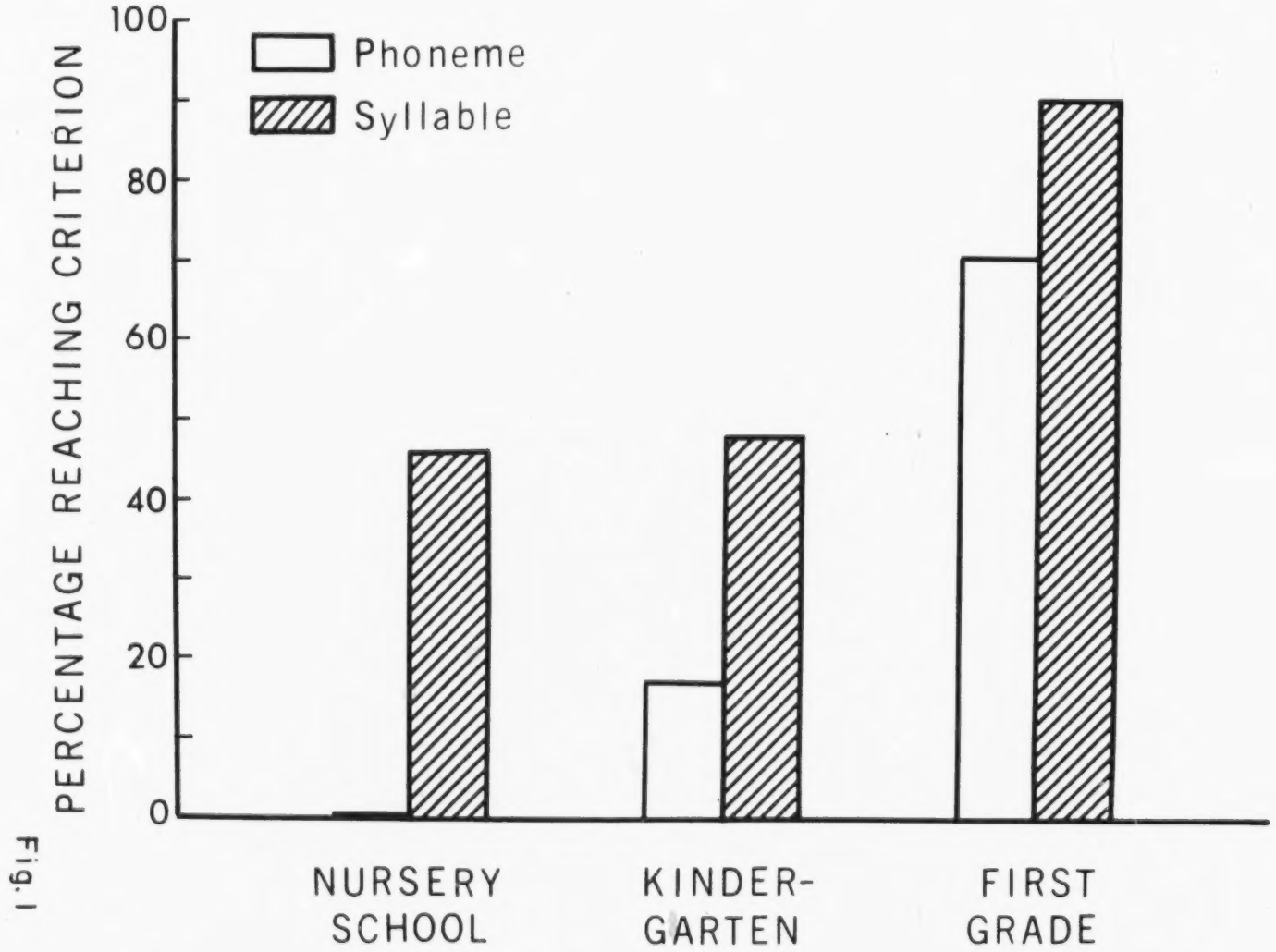


Fig. 1

Figure 1: Percentages of children reaching criterion in the phoneme and syllable groups at three grade levels.

The superiority of first grade girls on both tasks accords with the many indications of the more rapid development of language in girls (McCarthy, 1966). One might have expected, however, to find manifest superiority of girls at the earlier ages as well. At all events, it is appropriate to mention in this context that boys far outnumber girls among cases of reading disability (Vernon, 1960; Thompson, 1969; Critchley, 1970).

Though phoneme segmentation was more difficult than syllable segmentation at all three age levels, the phoneme task did show improvement with age. We cannot judge from this experiment to what degree these measured increases represent maturational changes and to what extent they may reflect the effects of instruction in reading. We would guess that the sharp increase from 17% at age five to 70% at age six in children reaching criterion level is probably largely a consequence of the intensive concentration on reading instruction in the first grade. To be sure, a certain level of intellectual maturity is undoubtedly necessary to achieve the ability to analyze words into phonemes. But there is no reason to believe that the awareness of phoneme segmentation appears spontaneously when a certain maturational level is reached. (If it did, we should think that alphabets might have been invented more frequently and earlier.) In any case, the possibility that these changes with age are relatively independent of instruction could be tested by a developmental study in a language community such as the Chinese where the orthographic unit is the word and where reading instruction does not demand the kind of phonemic analysis needed in an alphabetic system.

We are especially concerned to know more about those substantial numbers of first graders, some 30% in our sample, who apparently do not acquire phoneme segmentation. It would, of course, be of primary interest to us to know whether they show deficiencies in reading acquisition as well. It remains to be seen in further research whether inability to indicate the number of phoneme constituents of spoken words is, in fact, associated with reading difficulties. Our test can provide a measure by which differences in segmentation ability can be assessed directly. If we should find that performance on a test like ours can differentiate good from poor readers (let us say, among second and third graders), we should be encouraged to assume that inability to analyze words into phonemes is indeed a factor in reading disability. In any event, we would especially wish to determine whether more explicit instruction in phoneme segmentation by an extension of this procedure would be helpful in improving the reading ability of beginning readers.

We have here supposed that fairly explicit awareness of phoneme segmentation is necessary if the child is to discover the phonologic message and, ultimately, the meaning it conveys. But this is only a part, albeit an essential one, of a broader requirement: the orthographic representations must make contact with the linguistic system that already exists in the child when instruction in reading begins. Accordingly, the explicit awareness of linguistic structure with which we have been concerned is not necessarily the only condition that must be met, though we believe it is an important one.

REFERENCES

Critchley, M. (1970) The Dyslexic Child. (London: William Heinemann).

Doehring, D. G. (1968) Patterns of Impairment in Specific Reading Disability. (Bloomington: Indiana University Press).

Fant, C. G. M. (1962) Descriptive analysis of the acoustic aspects of speech. Logos. 5, 3-17.

Gleb, I. J. (1963) A Study of Writing. (Chicago: University of Chicago Press).

Kolers, P. (1972) Experiments in reading. Sci. Am. 227 (13), 84-91.

Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, I. Y. (1971) Basic research in speech and lateralization of language: Some implications for reading disability. Bull. Orton Soc. 21, 71-87.

Liberman, I. Y., D. Shankweiler, C. Orlando, K. S. Harris, and F. B. Berti. (1971) Letter confusions and reversals of sequence in the beginning reader: Implications for Orton's theory of developmental dyslexia. Cortex 7, 127-142.

Lindquist, E. F. (1940) Statistical Analysis in Educational Research. (New York: Houghton Mifflin).

Makita, K. (1968) Rarity of reading disability in Japanese children. Amer. J. Orthopsychiat. 38 (4), 599-614.

Mathews, M. M. (1966) Teaching to Read. (Chicago: University of Chicago Press).

Mattingly, I. G. (1972) Reading, the linguistic process and linguistic awareness. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 133-147.

McCarthy, D. (1966) Language development in children. In Manual of Child Psychology, 2nd ed., ed. by L. Carmichael. (New York: John Wiley & Sons) 492-630.

Monroe, M. (1932) Children Who Cannot Read. (Chicago: University of Chicago Press).

Rozin, P., S. Poritsky, and R. Sotsky. (1971) American children with reading problems can easily learn to read English represented by Chinese characters. Science 171, 1264-1267.

Savin, H. (1972) What the child knows about speech when he starts to learn to read. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 319-326.

Shankweiler, D. (1964) Developmental dyslexia: A critique and review of recent evidence. Cortex 1, 53-62.

Shankweiler, D. and I. Y. Liberman. (1972) Misreading: A search for causes. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 293-317.

Stevens, K. N. (1972) Segments, features, and analysis by synthesis. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 47-55.

Thompson, L. (1969) Reading Disability, 2nd ed. (Springfield, Illinois: Charles C. Thomas).

Vernon, M. D. (1960) Reading and its Difficulties. (New York: Cambridge University Press).

Machines and Speech*

Franklin S. Cooper
Haskins Laboratories, New Haven

INTRODUCTION

Speech, as a topic for the Conference on Research Trends in Computational Linguistics, was included in the agenda because of its emerging relevance to the field. This factor set the tone for the opening discussions within the group and was largely responsible for the fact that most of the group's attention was spent on the two topics for which the overlap with the subject matter of conventional computational linguistics was most clearly evident. There was tacit agreement that many of the problems that concern speech researchers and for which they regularly use computers would not be of interest to most of the conferees. Thus, there was little discussion of such current speech areas as the physiology of speech production and audition, perception of speech and speech-like stimuli, cross-linguistic studies of speech and universal phonological theory, children's language acquisition, speech and hearing disabilities, etc.

Instead, the discussion centered on speech production by computers in the context of reading machines for the blind and on speech recognition by computers as a central problem in designing speech understanding systems. There was in addition some discussion of the research tools needed both for research in depth at major research centers and for graduate level training in a larger number of university laboratories. The Chairman, in his report to the plenary session, dealt only briefly with the state of the art in speech research but put primary emphasis on research opportunities, covering some areas in addition to those which had received attention in the group discussions.

Speech as a Part of Computational Linguistics

Computational linguistics, as defined by past usage, has dealt mostly with written rather than spoken language. This is mainly an historical accident; nevertheless, the time has come to examine the areas in which speech may now be considered a part of this field. There are, indeed, new factors to be considered: one is that the domain of automated language processing is, in practice, being extended to include systems that generate speech as an output and that

---

*Report of discussions by a workshop group on Machines and Speech, held as part of a Conference on Research Trends in Computational Linguistics, Washington, D. C., 14-16 March 1972. Proceedings of the Conference, including this report, were issued by the Center for Applied Linguistics, Arlington, Va., 1972.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

accept it as an input. Another reason, based in theory, is a growing awareness of parallels between low level processes of speech production and perception and higher level processes involving syntax and semantics. This has come about mainly as a consequence of psycholinguistic experimentation on human language processing at all levels, although most of the effort--and the more penetrating methodologies--have been applied at the level of speech. Thus, research on human processing of language, speech included, can serve to suggest areas and methods that computational linguists may wish to explore.

This is not to say that all of speech research ought to be co-opted into computational linguistics: much that is merely descriptive or that is concerned with the technology of voice communication has little to offer or to gain in return. A reasonable criterion might be that there should exist a mutual interdependence between the speech processes and the higher level processes involved in the same overall automated language operation. This, at least, was the basis we used in choosing topics for discussion, and in considering practical applications. The nature of the interdependence will become evident in the following discussions of reading machines and of speech understanding systems.

Speech as an Output of Automated Language Processing: Reading Machines for the Blind

Most of the familiar instances of speech from machines--telephone directory assistance, airline announcements, etc.--are essentially uninteresting to computational linguists. They involve such limited vocabularies and fully defined syntactic structures that adequate speech can be had simply by using prerecorded words and phrases or their synthetic equivalents. The more general problem of generating speech from unrestricted text is only now approaching solution, primarily in connection with reading servic for the blind. Indeed, many of the problems that face any automated speech output device of a non-trival design can be described and analyzed by detailed consideration of this single application. The practical problem of providing such a service has many additional aspects, nonlinguistic as well as linguistic. Here we shall undertake no more than a sketch of component operations and problem areas.

The primary function of a reading machine for the blind[1] (so called because of the many attempts to build simple, portable devices) is to provide blind people with adequate access to the full range of printed materials that sighted people read. An approach that has been tried repeatedly during the past sixty years is to use a photoelectric probe that converts letter shapes into sound shapes which the blind user is expected to learn. This task, it is argued, is no worse than learning a foreign language; indeed, it should be easier, since the lexicon and syntax are those of English. In practice, learning is laborious and reading rates are disappointingly low--comparable to Morse Code and roughly an order of magnitude less than listening to spoken language. The reason for the superior performance of spoken language is by now quite clear: speech is a

[1] Two recent reviews that deal with the general problems of sensory aids for the blind (including reading machines) and that give extensive references to the literature are by Allen (1971) and Nye and Bliss (1970).

highly condensed encoding of the message, whereas the letter sounds provide only a serial encipherment into acoustic form and so are far less efficient carriers of linguistic information.

Both theory and experience suggest, therefore, that high performance by a reading machine requires that it be able to speak plain English--or whatever language is being used. The complexity and size of such a machine are such that it will have to be, for the present at least, a central facility that records tapes on request, or possibly serves a remote user who is linked to it by telephone. Many of the practical problems are therefore those inherent in the organization, financing, and administration of any sizable service function such, for example, as a time-sharing computer system.

In the operation of a reading service center, the first step is to obtain a machine-readable alphanumeric tape from the printed text that was requested by the blind user. Sometimes compositor's tapes (from which the book was printed) will be available, but usually the printed page must be converted by optical character recognition (OCR) machines, or by manual keyboarding. [This problem of entering data from the printed page is shared with many projects in computational linguistics. At present, there are no suitable OCR machines available as standard equipment, and only a few companies are prepared to supply special machines capable of reading proportionally-spaced characterers in multiple fonts. Service bureau facilities for reading conventional printed text are likewise limited and expensive. The reason for this situation lies only partly in the technical difficulties; mainly, there is less customer demand for such devices than for simpler and cheaper machines designed to read at high rates the specialized, uniformly-spaced characters from credit cards and business documents. It may well be that the needs of computational linguists, as well as those concerned with reading centers for the blind, can only be met through a development project aimed at their special needs, i.e., for moderate accuracy, moderate speed, and reasonable cost in a machine with enough virtuosity to recognize the commonly used fonts and to scan bound volumes. Whether or not such a project falls within the scope of computational linguistics remains open to question. It could nevertheless be so useful for text input to computers that a good case can be made for it.]

The next step, once the text is available to the computer in alphanumeric code, is to arrive at the pronunciation of each word in terms of some appropriate phonetic notation. Due to the nature of English orthography, no simple set of rules can be employed to derive an acceptable guess at the proper transcription for all English words, though there are spelling regularities that may, perhaps, be exploited to advantage. Hence, some kind of pronouncing dictionary (in machine-readable form) is essential. Two general approaches are being tried: (1) There is the straight-forward, pragmatic one of storing the phonetic equivalents for every word of a large lexicon, including separate entries for most root-plus-affix combinations. This allows the specification of inherent lexical stress, as well as of a code indicating the usual syntactic role of the word. A dictionary of this kind with approximately 150,000 entries, about equivalent to a desk-type dictionary, is easily accommodated on a single IBM 1316 disk pack. (2) With a more analytic approach, considerable savings in dictionary size, perhaps ten-to-one for very large dictionaries, may be achieved by attempting to break the input words into their constituent morphs. However, word pronunciation is not a simple function of the pronunciation of its constituent parts: for example, a suffix may shift the placement of primary stress

and force a change in vowel quality. In any case, the dictionary must contain phonetic equivalents for the full set of morphs and for a substantial number of frequent words that violate spelling-to-pronunciation rules. Research problems of considerable interest are involved in such an approach: for example, the development of a set of pronunciation rules based on an underlying representation of English morphemes, working out algorithms for word decomposition, finding rules for the placement of lexical stress in words that are reconstituted from their morphs, and inferring syntactic roles that such words can play.

Once a canonical phonetic representation has been obtained for each word in the text, it is necessary to employ a set of phonological rules to determine how the sequence of words which constitute an integrated sentence should be spoken. Perceptual experiments with concatenated words, using recordings of single words spoken in isolation, have demonstrated the importance and the extent of the sentence-level recoding of spoken word strings. As an example at the segmental level, a rule transforms the phonetic sequence [d, word boundary, y] into [ǰ], which means that the normal segmental realization of the sentence "Did you go?" is actually [dIǰugo]. While effects of this sort are familiar, the exact form which the rules take is not known and the inventory of such segmental rules is far from complete. Finding these rules is a challenging research problem and one to which computational methods can make an important contribution, especially in testing the reliability and range of application of proposed rules.

Of more importance than such segmental rules are phonological rules that determine the temporal organization of the acoustic output and the fundamental frequency of vocal cord vibration as a function of time. Segmental durations and intonation contours are influenced by a number of factors, especially the syntactic structure and stress pattern of the sentence. Indeed, some sort of syntactic analysis is essential for the synthesis of a satisfactory, spoken sentence. In some degree, structure can be inferred from orthographic punctuation, and rules based on punctuation are sometimes sufficient. Nevertheless, better methods are needed. This dependence of speech quality on syntactic structure is perhaps the major area of overlap between the reading machine problem and conventional computational linguistics, and a promising target for further research.

At this point in the process of generating speech from written text, the computer has assembled a phonetic string that has had appropriate allophonic and stress adjustments and that has been marked for intonation and juncture. It remains to convert this phonetic description into control signals that will operate a hardware synthesizer, or its simulation in software. Two general methods of speech synthesis by rule are currently utilized: (1) For a terminal-analog synthesizer, the rules manipulate acoustic variables directly, e.g., sound source type and pitch, formant frequencies, and intensities. In this case, the rules for synthesis begin by consulting tables for the canonical form of each phone, then computing the necessary contextual adjustments (corresponding approximately to coarticulation in human speech). Typically, there are about a dozen parameters that are used to control a hardware synthesizer; these are specified at regular intervals of about ten milliseconds, requiring a total output bit rate of about 4800 per second. (2) For an articulatory-analog synthesizer, the rules typically manipulate articulatory variables such as the positions and shapes of simplified models of tongue, lips, velum, larynx, etc. The resulting shapes (of the model vocal tract) are then used to compute an acoustic

output, or the control signals for terminal-analog hardware. (It should be pointed out that substantially more work has been done on terminal-analog models and rules which currently provide a more intelligible output than do articulatory implementations; however, articulatory models are improving rapidly and are of greater theoretical interest because rules of coarticulation may ultimately be built in as automatic constraints. In addition, modelling of this kind may lead to a better understanding of the physiology of speech production; indeed, this is an area of promise for future research to which the present discussion will return.)

The intelligibility of the synthetic speech currently produced by rule with terminal-analog systems is surprisingly good. Tests of consonant-vowel words and nonsense syllables synthesized by rule have shown that listeners are able to identify them correctly 95% of the time. Systematic testing for longer words and sentences has not been done, but interviews and informal tests have been carried out with blind students who have listened to chapter-length recordings of textbook materials. Few words are missed and overall comprehension is high. Nevertheless, it is clear that the treatment of consonant clusters and of stress and phrasing needs improvement and often seems unnatural. Future improvements in the rules will depend on careful analyses of natural speech and the systematic manipulation of synthetic speech to test the perceptual relevance of proposed changes in the rules.

The evaluation of synthetic speech in terms of its real-life utility to blind individuals is a major task. The speech itself needs to be considered along two dimensions--intelligibility and naturalness. The voice quality of current synthesizers is not as natural as would be desirable, despite efforts to improve it. This may imply that we are ignoring some critical variables such as glottal source irregularities or, alternatively, that naturalness will not be achieved until the synthesis rules are improved sufficiently to avoid all of the conflicts between acoustic cues and message content. It is interesting to note--and ultimately encouraging--that it is the rules in their present form and not the synthesis hardware that is at fault. In fact, extremely good synthetic speech--indistinguishable from the spoken version--can be made by meticulous hand adjustment of the control parameters.

But complete naturalness may be too much to expect for speech that is synthesized by fully automatic methods. Evaluation must then deal with the question of how useful the product is for, say, the blind student in preparing his lesson assignments. Its principal advantage over natural speech is that he should be able to get the material he wants when he needs it, not some weeks or months later as often happens with recordings by volunteers. The computer can read tirelessly and faster than a human, though its actual performance will depend on how well the service function is organized. Good intelligibility of the synthetic speech is, of course, essential but this may not be an adequate criterion. It might be that listening to synthetic speech imposes so much perceptual load that comprehending and remembering the content would be excessively difficult; hence, comprehension tests and measures of fatigue are more likely to be relevant than intelligibility tests in evaluating the practical usefulness of computer-generated speech. Evaluations of this kind are being started, in parallel with efforts to make the synthetic speech sound more natural.

This sketch of the reading machine problem has pointed to some of the areas of interdependence between speech research and computational linguistics. Thus,

many problems of dictionary management are shared. Spelling-to-sound rules and algorithms for decomposing words into constituent morphs would reduce the size of the dictionary needed for a reading machine, just as comparable algorithms for syllabification do for automated typesetting. Likewise, reliable methods for reconstituting words and for inferring their usage would be useful to either a machine that must read them aloud or to one that is composing written responses. But the common ground is most evident at the level of syntactic analysis. An efficient general purpose parser is almost equally necessary for properly rendering a sentence into spoken form, or for inferring its content from its written form. For the present, reading machines must depend on explicit punctuation and a pseudo parse of some kind; perhaps, short-cut methods that yield good speech would also have practical application to the automatic punctuation of synthetic written responses. The existence of this common ground implies research opportunities on the interrelations of spoken and written language--another topic for later discussion. But before dealing with areas for research and practical application, the report will give an account of the discussions on speech understanding systems.

Speech as an Input to Automated Language Processing: Speech Understanding Systems

Most of the discussions in this session centered on Advanced Research Projects Agency (ARPA)-sponsored projects on speech understanding systems, underway for about six months. Five major research groups, already engaged in other ARPA-supported work, are involved. As an initial step, a study group assembled from these projects analyzed the problems and prepared a set of objectives, specifications, and plans (Computer Science Group, Carnegie-Mellon Univ., 1971).

The overall objective is to develop one or more demonstration systems, primarily to show that the technology now exists--though it is scattered--to make such an undertaking feasible. Each of the major research groups is undertaking a task of its own choosing, but with the expectation that a cooperative effort on some one of these projects, or an entirely different one, will emerge. Definition of the tasks was considered to be a crucial point in setting the level of difficulty and the chances for success. Indeed, the long background of failures to solve the "speech recognition problem" was in part responsible for the ARPA decision to undertake a related, but more manageable, program.

The machine recognition of speech has been a persistent challenge for at least the past twenty years. Several ways have been found to recognize spoken digits, or even a few dozen words when they are spoken singly and carefully. But the general problem, usually put in terms of a phonetic typewriter or a system for automatic dictation, has remained elusive. The much larger vocabulary that is required, the necessity of dealing with connected speech, and the need to accommodate a number of different speakers have all posed severe difficulties. Moreover, as more has been learned about the nature of the speech signal and its understanding by humans, the clearer has become the magnitude and complexity of the recognition problem. It should be noted that most attempts thus far to deal with the general problem have used a "bottom-up" approach, i.e., one in which phonetic elements, words, and sentences, are all found by successive analyses based on the acoustic signal. The speech understanding systems projected by the ARPA Study Group differ in two important ways: constrained objectives make the problem more manageable, and reliance is not placed on bottom-up analyses.

Speech understanding, in the present context, means essentially that a computer, when told to do something, will "understand" well enough to take the correct action or will ask for clarification. If, for example, the computer has control of a robot, then a command to put the red block in the box can lead to an easily confirmed performance, provided there is a red block, and a box that is larger than the block; otherwise, it should lead to a question or an error message. Or again, if the computer contains a file of information about moon rocks, it should be able to answer inquiries about the numbers, sizes, and chemical compositions of those rocks, whether the query is phrased as a question or as a demand for data. Obviously, tasks of this kind are multidimensional and can be constrained in ways that will, in fact, determine their difficulty. The attempt has been to choose tasks--more or less like the two mentioned--that are constrained in such a way as to make them manageable, but not to the point of making them trivial. A practical payoff was not considered a mandatory requirement.

Performance in the task situation not only limits the number of possible responses to the voice input, it also provides additional bases for analyzing it. Thus, limitations on the vocabulary, on the syntax that is permitted for the task, on allowable operations to be performed, and predictions of the probable behavior of the person speaking--all provide bases for making hypotheses about the actual massage carried by the incoming acoustic signal. Indeed, it is on such lexical, syntactic, semantic, and pragmatic "support" that the research groups are putting their hopes and focusing their efforts; although the acoustic signal is not being ignored, it is receiving far less attention than it has in past efforts to solve the general recognition problem--quite possibly, less attention than it will need to receive if effective use is to be made of "top-down" or "side-in" approaches.

The ARPA Study Group's final report discusses at length the kind of support, and the nature of the problems, to be expected at the various levels. Interdependence across levels is a characteristic of the entire undertaking, and there is much overlap with the domain of conventional computational linguistics. There are challenging differences, too: neither sentences nor words are well formed, as one would expect them to be in written text; moreover, the need for live interaction between user and computer means that the computations must be done in real time or less, much of it while the sentence is still being spoken.

Only at the very lowest levels, where parameters are being extracted from the acoustic signal and are used to guess the phoneme string, are the problems wholly within the domain of speech research. Here, although engineering problems of pitch derivation and formant tracking are not trivial, the major difficulty lies in inferring the phonetic string--a difficulty that may be inherent in the nature of speech itself. It is clear that in the production of speech there is much restructuring (or encoding) of the segmental units to achieve a compact, smoothly flowing output; that is, there is much overlap of the gestures for units that are themselves sequential. It is, indeed, the rules for this coarticulation that provide a basis for speech synthesis by rule. But the rules we use are generative rules; the inverse rules, to the extent they exist, are largely unknown except for those phones and contexts where coarticulation is minimal and the "code" can be said to be transparent. A general paradigm, proposed by Stevens (1960) some years ago and labelled "analysis-by-synthesis," uses heuristics to guess the phonetic string and then confirms by synthesis, or uses error signals to make a better guess. An alternative,

referred to by the Study Group as "hypothesize-and-test," looks for transparent places in the code (and for other acoustic landmarks) to generate a much less complete hypothesis about the phonetic string, and then brings to bear information from higher level processes. An example would be to look for a word that begins with strong, high frequency noise, then guess the word to be "six" if it has one syllable, or "seven" if two, provided one has other reasons to expect a spoken digit.

Obviously, there are significant research problems at the speech level, as well as practical difficulties, for a speech understanding system. Can a set of rules for analysis be devised? Or can even a partial set having practical utility be devised? What kinds of heuristics will be most helpful if one resorts to analysis-by-synthesis? What relative reliance should be placed on extracting as much information as possible from the acoustic signal versus having to depend on support from higher linguistic levels? Underlying these questions is the assumption that present knowledge of the acoustic cues is reasonably complete; in fact, much remains to be learned about the cues that operate in consonant clusters and in connected speech, both careful and casual.

At the next higher level, the principal problem is how to convert the string of phonetic elements--perhaps incomplete, and certainly error-prone--into a string of words, which may also have intervening gaps. Word boundaries are not at all evident at the acoustic level, and so do not appear in the phonetic string. Segmentation into lexical elements must then proceed mainly by matching to strings that correspond to words in the task vocabulary. Often the low-level analysis will have generated more than one possible candidate for each phone, omitted an element, or supplied a wrong entry. The matching operation must somehow avoid the combinatorial explosion that could readily occur if there were several options for each of a string of phones. Omissions and errors pose obvious additional difficulties. There is, therefore, serious need for support from higher levels as well as efficient analysis at the phonetic level.

The construction of sentences from a partial and "noisy" string of words can draw support from whatever restrictions the task may impose on the syntax, or from information about the location of syntactic boundaries that can be inferred from the suprasegmentals found in the acoustic signal. The latter relationship is essentially the inverse of the dependence of synthesis on syntactic information to assign stress and intonation and thereby generate speech that sounds natural. There are other problems at the sentence level that have their counterparts in parsing written text. They differ, though, in that analysis of the spoken sentence will often have to deal with intrusive hesitation sounds, with non-well-formed or incomplete sentences, and with both backwards and forwards parsing from a starting point somewhere within the utterance. There are obvious research problems of considerable importance in devising methods for parsing under these conditions.

Interpretation of the sentence as a basis for action can rely only in part on the output of the syntactic analysis, since that output is likely to be faulty or ambiguous. Good use will need to be made of the semantic constraints imposed by the task, and of pragmatic information about the behavior to be expected of the human operator. Except for such help, the task of interpretation has all the complexities inherent in the content analysis of written text. If the appropriate response from the computer is an answer to a user's question, then various cognitive functions must also be performed and a sentence must be

generated that is appropriate in both content and form. The response itself could be in either written or spoken form, with the latter making use of techniques for text-to-speech conversion developed for reading services for the blind.

It will be evident that speech understanding systems are involved with language at all levels, and pose many problems that should interest computational linguists. The differences between these problems and more familiar ones reflect deep differences between oral and written language. Even so, this account has probably understressed the pervasive influence of speech at all levels of the speech understanding problem, and not merely those levels to which the term is usually applied.

Research Areas With Speech Involvement

The preceding discussions of speech outputs from computers (for reading machines) and speech inputs to computers (in speech understanding systems) have exposed a number of important areas for research in computational linguistics. Some of the basic problems that were mentioned in passing deserve additional consideration.

Translating between written and oral language. Although it may be overdramatizing the differences between written and oral language to speak of translating the one into the other, it may nevertheless suggest a useful point of view in reexamining some old problems and considering some new ones. We have seen that the research on reading machines for the blind makes explicit some written/oral differences that often pass unnoticed. Ideally, a reading machine should convey all of the useful information that is on the printed page. This may include much more than the bare text we have so far considered, even without taking account of pictures and diagrams, a task far beyond present capabilities. The well printed page uses many typographic devices to organize and modify the literal text: punctuation marks are so commonly used that we think of them as a part of the text, though their realization in sound follows very different rules than those applied to the letter; capitalization, too, an additional aid is widely used, and for a variety of purposes; an additional aid is the judicious use of different type fonts, sizes, weights, and leadings; and finally, paragraph identation is another of the devices used to indicate breaks, subordinations, listings, etc.

All of these carry valuable information to the eye. How, and to what extent, can this information be "translated" for easy use by the ear? Not, one would hope, by overt description of the typography. The commonest forms of punctuation have acoustic reflexes that are fairly simple and regular. Are there comparable acoustic signals for other graphic symbols? The inverse transformations involved in writing are no easier, though they are more familiar: thus, oral questions, statements, and exclamations are indicated by their standard typographic signs; likewise, emphatic stress can be signalled by italics. But how does one convey a note of incredulity or petulance without resorting to bald description? We know that the skillful writer can do it; perhaps, then, ways can be found to convey typographic messages of comparable subtlety to the blind listener in a reasonably graceful way. As a practical matter, when one is converting a compositor's tape into speech, something must be done with each of the graphic signals that it contains.

If compositors' tapes seem to carry too much information about graphic details, they can be accused equally of carrying too little of the essential information about the sentence. Where, for example, is sentence stress to be placed during synthesis? Where are the breaks into breath groups for a long phrase? And just how should the intonation be handled? But it is hardly correct to say that the information--or most of it--is not provided; rather, it is buried in the structure, which is why good speech synthesis will be so dependent on adequate parsing. This assumes though that full knowledge of structure is a sufficient condition for making speech that sounds natural. Even if true, much has yet to be done about finding the rules by which sentence structure can be converted into sentence suprasegmentals. Rules for the inverse transformation from suprasegmentals to structure--to the extent that they exist--could be extremely helpful in supplementing the phonetic analysis in a speech understanding system. A related problem that dips more deeply into conventional speech research is the relationship between stress and intonation (as linguistic entities) and the relevant dimensions of the acoustic signal. These relationships are known only in part.

The written sentence is, nevertheless, incomplete in at least some respects: witness sentences that are ambiguous in written form but not in spoken form, because the real-world context is missing in the one case but often is indicated (by stress and intonation) in the other. One can find ambiguity at the lexical level, too--witness homographs and homophones--though the ambiguity in one mode of expression is usually resolved in the other. It might be interesting to explore the conditions for ambiguity wherever it occurs.

These are a few of the translation problems that one would encounter in dealing only with English. It is easy to see, from the different structures and orthographies of other languages that different--but probably not fewer--problems would arise with them. Finally, the general problem of working back and forth between oral and written versions of the same language will be one of increasing concern to computational linguists as automated language processing becomes more and more involved with speech as its input and output modes.

Modelling speech production. We have seen that the rules by which speech is synthesized necessarily deal with the higher level processes that are the normal domain of computational linguistics. Although the models of speech production to which those rules apply lie more nearly in the usual realm of speech research, the models must operate from higher level control signals. Hence, the development of speech models and the organization of their control signals is an area of research that is relevant to computational linguistics as well as important in itself. Additional reasons are that the processes of speech production parallel those at higher levels in interesting ways, and that experimental methods for probing the lower level processes are well developed.

The process of human speech production has several distinguishable subprocesses, organized hierarchically into levels. Parallels with the levels of linguistic processing are based mainly on operations that restructure the intended message to make it more compact and to put it into linear form, an obvious requirement for eventual output as a time-ordered acoustic signal. Speech has the additional feature that its processes change mechanisms on the way down; implementation is no longer done at all levels by neuromechanisms, but must include signal transformations in proceeding from nerves to muscles to gross movements and to sound generation. Thus, some of the restructuring, or encoding, that we

find in the spoken message is a consequence of interface requirements. Linguists usually stop--perhaps with good reason--when they have specified the phonetic string (or its equivalent in the form of a sequence of feature matrices), leaving actualization of the message to a human speaker.

What remains to be performed--and to be modelled--are the successive conversions that lead eventually to speech: (1) The phonetic string must be grouped into pronounceable units and converted into a pattern of neural commands to muscles. This involves collapsing the string of linguistically discrete elements into an articulatory unit of about syllabic size and the temporal coordination of a substantial number of neural signals. Some of them may be crucial for maintaining phonemic distinctions (and some merely accessory), or all may be directed at achieving a particular target performance. The organization of these unit gestures is a topic of very lively interest in current speech research. (2) The pattern of neural impulses activates the muscles of articulation in a process that may be quite straightforward, or may involve gamma-efferent feedback loops in important ways--another topic of current interest. (3) In either case, the gesture in neural form is converted, subject to muscular and mechanical restraints, into gross movements of the several articulators, and these in turn determine the configuration of the vocal tract and its acoustic excitation. This involves encodings of quite a different kind from those mentioned above, but often quite extensive; they account at least for a mechanical component in coarticulation. (4) Finally, excitation and configuration determine uniquely--but not simply--the acoustic waveform of speech.

Efforts to model these processes have typically worked upstream from the acoustic level, usually dealing with a single conversion. Thus, the work of Fant (1960) and others has given us a good grasp of how to convert changing articulatory shapes and excitations into changing acoustic spectra. X-ray movies and sound spectrograms are only two of the experimental methods for exploring and testing these conversions. The relationships between muscle contractions and articulatory movements are under intensive investigation, using electromyography both to measure muscle activity and to infer neural signals, and using x-rays and spectrograms to observe and infer the resulting movements. Efforts are being made to describe the organization of gestures in motor command terms, with verification to be provided by measurements on muscle activity, configurations, and sound.

Computer methods have been used to good effect in both the experimental work and in modelling conversions at the lower levels. They have been used also to good effect, but in quite different ways, in speech synthesis by rule. Thus, the terminal-analog type of synthesis by rule bypasses all the intermediate stages and operates directly from an input phonetic string to the output speech waveform. The articulatory type of synthesis by rule makes a lesser leap from phonetic string to articulatory gestures, then uses level-by-level models to get to the acoustic output. The obvious goal is good modelling of the conversions at <u>each</u> level, confirmed by direct experimental measures wherever that is possible, and also by the synthesis of natural-sounding speech when the models are used in tandem.

<u>Interfacing speech to phonology</u>. It might appear from the preceding discussion that the processes of speech begin where those of linguistics usually end, i.e., with the message in the form of a phonetic string or the equivalent sequence of feature matrices. However, the phones and features of the speech

researcher are not--or not necessarily--those of the linguist, since they are defined by different operations. It is the resulting mismatch at this level that poses one of the major problems in modelling the total process of generating spoken language.

When linguists use labels such as "compact" and "diffuse" for features, there is no implication that the features will convert directly into components of a pattern of neural commands to muscles; when such labels as "voiced" and "voiceless" are used, the differences in operational definition are concealed, though they are no less real. The differences have their origins in the dissimilar approaches taken by linguists and speech researchers. The latter usually try to work with real mechanisms and models of processes whereas linguists more often concern themselves with relationships that they can formalize as rules, though exceptions are to be found on both sides.

The interface problem, then, has two different complexions. If linguists' rules really reflect underlying mechanisms and processes--a claim that linguists rarely make for them--and if current speech models prove to be tenable then a conversion is surely possible and finding it becomes an important research goal. But it is conceivable that the linguists' rules are wide of the mark as to processes, however useful they may be as descriptive devices. It would not then be possible to find the "real" conversion, though the search for it might make clear the directions in which phonological theory ought to move. In any case, the problem is inescapable in some guise when automated language processing must operate across the boundary between speech and phonology.

Converting generative rules into analytic rules. The discussion of modelling speech production, including the special case of interfacing speech to phonology, has all been generative and most of the models that are concrete and believable are likewise models of production processes. This does not imply that perception has been less studied than production, but only that the research has yielded a more coherent set of relationships for the latter.

There may be a good theoretical reason why this is so: the production process includes important operations that are in principle irreversible--irreversible in the same sense that a drainage system would be irreversible, i.e., water does not run uphill and, if it did, it would not know which way to go at the confluence of two "downhill" streams. To the extent that speech perception is organized in motor terms and shares these irreversible operations, it cannot be expected to provide a model for straightforward analytic rules. Put another way, the production of speech involves encoding operations and so one must expect that the inverse operations, like decoding in cryptography, will be inherently complex and liable to ambiguity.

An alternative view of speech perception does not link it to the motor system and so evades any need to run that machinery backward. It puts its dependence on auditory mechanisms, starting with feature detectors, and employs processes that are in principle describable by models and analytic rules, though these have yet to be discovered.

Clearly, the nature of speech perception is a central problem for speech research. Its relevance to computational linguistics, already discussed in connection with speech understanding systems, lies in how it affects one's choice of strategy in choosing methods for inferring the phonetic string from

acoustic parameters, i.e., whether to stress analysis by synthesis with all its inherent difficulties or to concentrate on finding analytic rules, accepting the risk that they may not exist in any useful form.

Applications With Speech Involvement

One can be reasonably certain that the practical applications of automated language processing will not lag far behind the development of a technical capability.[2] It is easier to foresee examples that involve written-to-oral conversions than oral-to-written, so they will be discussed first.

Reading machines. Synthetic speech as a reading service for the blind has already been discussed at length in terms of the research problems involved in setting up a central facility to make tape recordings from books. There is genuine need for such a service, especially for students and professionals. The practical objective is to have several service centers scattered across the country so that mailing delays will not be excessive. This will have to be preceded by a shakedown of the methods (still research oriented) and then some operating experience with a pilot center that uses production methods and equipment. It should be possible to accomplish both tasks within about five years, and to have begun the establishment of a network of service centers.

An obvious extension is to allow local users to have on-line access to the text reading facility and, as a second stage, to make this access available by telephone. The latter would pose formidable problems if character recognition were performed centrally and only text scanning were done remotely. It may be, however, that newer methods of feature extraction, or total recognition of the printed text, will have been developed by that time, and so would make the data transmission problem quite manageable. Nevertheless, real-time continuous processing poses very different problems from those of batch processing, some very similar to the problems encountered in real-time interaction with a speech understanding system.

Remote retrieval of information. The same technology that reads for the blind can be used to allow quick access to library holdings by telephone from a remote location. Many of the local requirements for such a service will be met for other reasons in any case, so the additional investment need not be large. Thus, some types of library holdings (abstracts, bibliographic information, etc.) are increasingly being supplied and stored on magnetic tape, with programs that provide fast access to desired items. With a little help from the reference librarian, machine-readable information could be found and processed by the library's computer to yield synthetic speech which the remote user could listen to by telephone. Such a service will not answer all needs, of course, but it should be valuable in many instances and it has the great virtues of requiring few additional central facilities and of being able to use the existing telephone network instead of special terminals.

---

[2] The current status of research and development in this area is reported in the Conference Record of the 1972 Conference on Speech Communication and Processing, April 24-26, 1972, at Newton, Mass. [The Record is available from the National Technical Information Service or the Defense Documentation Center, AD 742236.]

An obvious elaboration is to allow the human caller to speak directly with the computer--another application of speech understanding systems. Even without this complication, however, there are important linguistic problems involved in remote information retrieval. An obvious one is that much of the information now stored in machine-readable form is very "dense" in that it uses many abbreviations and graphic devices. Even connected text is often telegraphic in style. The reinterpretation of such information, now organized for the eye, to make it suitable for the ear becomes almost a condition for telephone access.

Computer assisted instruction (CAI). Most CAI terminals operate solely with visual output and keyboard input, not because these modes are always optimal but because other modes pose major technical difficulties. Speech output, in particular, would be highly desirable in many cases, and is clearly the method of choice for much of the interaction with children in the lower grades. They could benefit from a great deal of content instruction if it were presented orally, but they do not have the reading skills to cope with it visually. For older students, too, oral information would often provide a useful supplement to visual displays.

This enhanced capability for CAI requires little more than adaptation of the text-to-speech techniques developed for the blind; in fact, the problem of providing good speech is easier, since the instructional text can be stored as a marked phonetic transcription that has been hand tailored to give natural sounding speech. Moreover, the storage requirements--hence, the possibilities for truly interactive CAI programming--are essentially the same as for literal text. Thus, the real utility of synthetic speech to CAI is likely to be far more dependent on imaginative programming than on technical limitations in providing spoken responses.

The ideal arrangement, in adding a speech capability to CAI, would be to let the machine respond appropriately to spoken responses by the student. Special purpose solutions, comparable to digit recognition, might work very well in many cases, especially with older students. But the greater need, and certainly the greater technical challenge, lies in making it possible for the younger student to interact in a reasonably free manner with his automated instructor. Clearly, this involves all of the problems of speech understanding systems as currently envisaged, compounded with the technical problems of processing children's speech and the linguistic problems of dealing with their free-form syntax. As a practical matter, it would be a mistake to hold back on the use of speech as an output in the hope of an early solution to the input problem, despite the many advantages that two-way speech would have in enlivening the interchange and removing artificial constraints on instructional programming.

Voice typewriter. The prognosis for typing or typesetting under voice control is probably no better than that for voice input to CAI. It is apparent by hindsight that the choice of the voice typewriter as an initial target for research on speech recognition was a serious error. Such a machine must deal with unrestricted inputs and a wide range of speakers and dialects. One can scarcely imagine a practical task of greater difficulty! However, both the nature of the problem and paths to intermediate goals that might lead on to an eventual solution have become much clearer. Thus, on the one hand, what we know of the nature of speech tells us that pattern matching will never provide a general solution, no matter how sophisticated the techniques; on the other hand, the use

of a "side-in" approach to speech understanding problems of limited scope promises to be one of the paths that might eventually lead to general speech recognition.

Speech understanding systems. The nature of the problems, the difficulties to be expected, and some of the areas in which research in both computational linguistics and speech can be helpful have already been discussed. The tasks in which speech inputs are to be used were chosen as demonstrations rather than practical applications. Even so, they are difficult enough to pose very real research challenges.

The question is often raised about what, if any, really practical applications exist for voice input or, in more realistic form, what practical tasks there are that are not handled adequately by more conventional and less complex means. The ARPA Study Group listed some eight tasks as examples of practical applications: airline-guide information service, desk calculator (with voice input), air traffic controller, missile checkout (accepting spoken comments and questions from a human inspector), medical history taking, automatic protocol analysis, physical inventory taking (involving voice interaction with a human inspector), and robot management by voice. If none of these seems of compelling urgency, it may be in part a reflection of the fact that our capabilities in speech recognition are still so primitive as to shackle our imaginations.

Man-machine cooperation. In summary, there are several practical uses to which speech outputs from automated language processing can be put, and probably will be put in the near future, though the prospect for practical application of speech inputs seems more remote. But one is tempted to say about speech recognition that, like Everest, it is there--and eventually the challenge will be met.

Man-machine cooperation with computers is already a fact of life, though at present that cooperation can only be had on terms that are convenient to the computer. Again, one is tempted to say that such a state of affairs cannot continue; it is a safe prediction that man will insist on cooperation on his own terms. This means that computers must learn to listen as well as to talk. It will not matter much that this involves complexity and expense; if these were paramount considerations over the long term, we would all have telepraphs in our homes instead of telephones.

Instrumentation for Research and Training

The objective of the session's participants in their discussions of machines and speech, was to consider not only promising areas for research but also needs and new possibilities for research tools. One suggestion that met general approval was that a state-of-the-art survey be commissioned to discover what the needs really are and to make generally available a knowledge of recent developments in the leading research laboratories. Very often, new devices or software which are built to fill a local need do not seem to the investigators to be sufficiently important to justify separate publication. Hence, they remain unknown, except to a handful of visitors.

This led to a discussion of how widespread the need might be for sophisticated new research instrumentation. The need is, or course, dependent on the number of centers in which basic speech phenomena are being studied intensively,

and on the prospect for additional centers. On this basis, instrumentation needs are comparatively modest quantitatively, though crucial for the limited number of major research centers that do exist--perhaps half a dozen in the United States, and comparable numbers in Western Europe and in Japan. The establishment of additional centers is made difficult by the "critical mass" (for both men and equipment) that is needed to do effective research; indeed, the increasing complexity of adequate tools and the need for cross-disciplinary approaches seem likely to increase the pressures toward centralization of research. There are, on the other hand, both a need for well-trained people, and a number of good academic centers where training could be much improved by enough research equipment to make that training modern and realistic. The sound spectrograph is one such tool that is now rather widely available. Speech synthesizers, on the other hand, which could be at least as useful, are rather rare. This seems unfortunate, especially since the technology is well known and the costs are not excessive. Thus, a good background for many of the kinds of speech research described in preceding sections could be obtained with a mini-computer plus disc file, obtainable for $10-15,000. The point was made that the much larger computers already available for batch or time-shared use at many universities are not adequate substitutes for even a mini-computer that can be used on-line; in fact, very few computer systems can handle speech, primarily because of the high, continuous data rates that are required. A state-of-the-art review could be particularly useful to schools that wish to install a training facility of the kind described, not only in alerting them to the possibilities, but also in providing detailed information that often takes a great deal of time to learn by trial and error--a familiar experience, summarized by one discussant: "the first program costs a year."

Some of the new developments and trends to be expected in research instrumentation will be cheap mass memories of very large size and a new order of magnitude in central processor speeds. There are general consensus that the trend toward interactive systems that operate in real-time will continue, with a large payoff in research productivity. Likewise, new facilities for graphic output are becoming easily available and will be most useful.

In summary, although this part of the discussion found continuing progress and no urgent needs in the limited number of centers where most of the basic research on speech is done, it delineated a considerable need to upgrade awareness and training facilities in a much larger number of university centers in order to enable their graduates to become familiar with modern methods and problems. A state-of-the-art survey would be a useful initial step.

Conclusion

The group's discussions concerning Machines and Speech dealt mainly with the nature of the research problems that are encountered in incorporating speech into computational linguistics. Two specific applications--reading services for the blind and speech understanding systems--were discussed at length. Both are examples of an increasing trend toward automated language processing and, in particular, extensions of this technology to the use of speech as an input-output modality.

The group identified a number of specific areas in which there is strong interaction between the usual domains of speech and computational linguistics. Thus, for example, the synthesis of natural-sounding speech, starting with written

text, requires information about the placement of sentence stress, about durations, and about pitch contours. This is information that is implicit in the structure of written sentences; hence, good synthesis would seem to require a capability for parsing. Conversely, in attempting to infer sentences in written form from an oral input, the suprasegmental information could provide much help in assigning structure to a string of phonetic elements. In general, many of the problems--and some very promising areas for research--lie in the conversions that must be made between language in written form and language in oral form. Thus, the addition of speech as input and output modes for automated language processing will necessarily focus attention on a whole set of problems that might otherwise pass unnoticed.

There are other areas of speech research that also interact with higher level processes. Thus, efforts to build detailed models of the processes of speech production (and to apply them to synthesis) must start with a description provided by phonology, and so cannot ignore the interface--presently missing--between speech processes and phonological rules.

Practical applications that make use of speech as an output from automated language processing are well on the way to being realized: reading services for the blind, remote retrieval of information by telephone, and a vocal response capability for computer assisted instruction. The prospects for processes that use speech as an input are more tenuous, though a major effort is under way to build demonstration models of speech understanding systems, i.e., computers that will accept instructions or questions via microphone. For the long term, there is little doubt that man-machine interaction will become increasingly important in a practical sense, or that there will be a steady pressure on the machine to conform to human convenience, i.e., to learn to talk and to listen.

The state of speech research here and abroad was also discussed briefly and it was noted that the trend toward concentration of research in only a few major centers is likely to continue because of the critical mass of men and instrumentation needed to deal with problems that are increasingly complex and multidisciplinary. But adequate research training need not be correspondingly concentrated; the provision of modest research facilities--in particular small computers used for synthesis studies--could do much to broaden the base of research training. A state-of-the-art survey and prospectus would be a useful first step.

REFERENCES

Allen, Jonathan. (1971) Electronic aids for the severely visually handicapped. In Critical Reviews in Bioengineering, ed. by D. Fleming. (Chemical Rubber Co.) 139-176.

Computer Science Group, Carnegie-Mellon University. (1971) Speech Understanding Systems: Final Report of a Study Group. (Pittsburgh, Pa.: Carnegie-Mellon University).

Conference on Speech Communication and Processing. (1972) Conference Record. (Conference held April 24-26, 1972 at Newton, Mass.). [NTIS or DOD, AD 742236]

Fant, C. G. M. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Nye, P. W. and J. D. Bliss. (1970) Sensory aids for the blind: A challenging problem with lessons for the future. Proc. IEEE 58, 1878-1898.

Stevens, K. N. (1960) Toward a model for speech recognition. J. Acoust. Soc. Amer. 32, 47-55.

An Automated Reading Service for the Blind*

J. Gaitenby, G. Sholes, T. Rand,[+] G. Kuhn,[+] P. Nye, and F. Cooper
Haskins Laboratories, New Haven

ABSTRACT

This is a progress report concerning the state of the reading machine that has been designed and developed at Haskins Laboratories, and that is about to be evaluated in field tests. After being exposed to standard comprehension tests and making judgments on rate preferences, blind students--performing as subjects in the testing--will assess the relative utility of synthetic speech recordings in comparison with face-to-face readings and naturally-spoken tapes.

INTRODUCTION

We are concerned with the problem of getting the results of research on a reading machine for the blind out of the laboratory and into application. The reading machine in question converts printed material to synthetic speech. It is hoped that within three to five years a pilot version of the machine can be installed in a university library to assess the feasibility of eventually constructing a larger-scale automated reading service for the blind--perhaps on a national basis. At present the methods for text conversion to artifical speech--and the speech itself--have reached an advanced stage of development. An optical character recognition device, forming the input of the system, will be received from a manufacturer within a few months. Meantime, editing of the "spelling and sound" dictionary (in which text word orthography is matched to phonetics) continues along with further refinement of the dictionary word retrieval and stress assignment programs. Modifications in the speech synthesis procedure are under way for improving the naturalness of the spoken output, and on a separate front, attention is being directed to the planning and arrangement of extended evaluation studies using texts generated in synthetic speech. This paper contains a short description of the steps used in the automated reading process and some of the plans for a full-scale evaluation of the synthetic speech output. (Accompanying the oral version of the paper were taped samples of synthetic speech illustrating the performance to be expected from a reading machine.)

---

*Presented at the 1972 Carnahan Conference on Electronic Prosthetics, Lexington, Ky., September 22, 1972, by P. W. Nye.

[+]Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

There will be very few, if any, blind people present who will dispute the statement that existing reading services which use tape recordings have shortcomings. These become particularly obvious when the services attempt to produce new tapes of recently published books. Sometimes months can elapse following a request before the recording is completed and the last chapter is received by the subscriber. The reasons for these delays do not, however, stem from glaring faults in the structure and efficiency of the organizations involved, but rather from the fact that they use human readers. Whether the reader is an actor working for a fee, or a volunteer, he or she must schedule visits to the recording studio when the facilities are available, and can then only work well for an hour or two at each session. Therefore, a major limitation lies with the human reader, and a solution to part of the problem appears to be available through the use of reading machines.

Techniques for recognizing printed characters and for synthesizing speech have been steadily improving over the past fifteen years. During this period the staff of Haskins Laboratories have been concentrating their efforts on the problems of speech synthesis and, with the support of the Veterans Administration, have been actively engaged in the development of a reading machine for the blind. At this point, sufficient progress has been made to make it obvious that a complete reading machine, which can produce speech from a printed page, can now be made. In fact, with the acquisition of an optical character recognition machine provided by the Seeing Eye Foundation, we expect to have a complete working model in the laboratory within the next few months. We have been able to generate long texts in the synthetic speech output for about two years, and stress and intonation assignment is now programmed in addition to the earlier speech synthesis by rule. (In the oral version of the paper a demonstration tape was played at this point.)

As you have noted, the speech is not perfectly natural and requires a little exposure to become used to. However, the words you heard were delivered at a final rate of over 160 words per minute and I am quite sure that if I were to ask you questions about the passage you would be able to provide answers to most of them. My confidence in your probable reactions is due to the fact that we have run pilot tests on comprehensibility of the material with college students, with blind veterans at the Eastern Rehabilitation Center of the West Haven Veterans Hospital, and with ourselves as subjects.

The results of these tests have been most encouraging and we feel that the reading machine system can be of real utility, particularly to blind students. However, in our efforts to apply the machine to student needs, we find ourselves obliged to take on several new problems, many of which lie outside the usual confines of a research laboratory. In fact, it is becoming clear that it will be necessary for us to conduct a thorough analysis of the uses to which synthetic speech can be put--and eventually to build at least one pilot reading service center before other agencies are likely to grasp the initiative. This means that we must continue to use our laboratory facilities for generating synthetic speech, and must conduct an extensive evaluation program in an endeavor to provide sufficient evidence to justify the investment needed to establish a Reading Service Center.

THE APPLICATION OF READING MACHINES

We anticipate that the Reading Service Center (in which the reading machine will operate) will be located in a library. At the Center printed texts will be

converted to synthetic speech and the outputs will be recorded for use by a large number of blind subscribers. The reading service can be provided in response either to a personal request made at the library or to a phone call. At the time his request is filed, the user may specify the word rate at which the material is to be recorded, as well as the book, article, chapter, or page he needs. Within a comparatively short time (minutes, hours, or possibly days, if request traffic is exceptionally heavy) he may pick up, or have sent to him, the audio tape of the entire text desired.

Because of the present limitations on the naturalness of synthetic speech, it is apparent that a reading service involving text-to-speech processing by machine can only supplement--and will certainly not supplant--existing reading services (Nye, 1972). Thus book production in Braille (as at the American Printing House for the Blind) and spoken tapes issued by Talking Books, Recording for the Blind, and other organizations, may long remain primary sources of reading material for the blind. However, in the educational field where there is a widespread need for more access, and more rapid access, to the printed word, it is probable that blind students will frequently accept a somewhat unusual voice output in exchange for an extremely fast supply of diverse published material that is not immediately available elsewhere.

THE TEXT-TO-SPEECH SYSTEM

The procedural steps in the reading machine system designed at Haskins Laboratories have been detailed in several other publications, and only a short review will be given here. After the printed text has been read into the machine, three successive transformations are involved: first, from English spelling to phonemic transcription of the same words; second, from a phonemic transcription to control signals; and third, from control signals to audible synthesized speech (Cooper, 1963). In somewhat more detail, following Nye et al. (1972) the text-to-speech processing is accomplished in the following way.

The requested text is scanned by an optical character reader (OCR). (The Cognitronics System 70 page reader, a machine that recognizes OCR-A upper and lower case typefont, is to be used for the scanning operation.) The OCR output, stored on digital magnetic tape, is transcribed to phonetic spelling with the aid of a 140,000 word text-to-phoneme dictionary that is stored in computer memory.

This stored dictionary, which was culled from the dictionary prepared by Dr. June Shoup and associates at the Speech Communication Research Laboratory, has been extensively revised to make it compatible with the local computer programming. The vocabulary has been separated into two units: a small high frequency word list is stored in core memory while the main lexicon is held in disc storage. Any text word encountered in the scanning stage is first searched in the high frequency list, and if not found, is then searched in the main unit. When the phonetic words of a sizable body of text, plus their grammatical categories, have been retrieved, a pseudo-grammatical comparison of successive paired words is made. Using a system of lexical and punctuational rules, stress symbols are assigned which are appropriate to the context (Gaitenby et al., 1972). The resulting prosodically-annotated string of phonemes (in machine language) is processed by the Mattingly (1968) program for Speech Synthesis by Rule, and control signals for the specific textual message at a predetermined speed are computed. These signals control the synthesizer and generate intelligible speech at the specified word rate. The output of the synthesizer is recorded on standard audio tape and conveyed to the blind user.

To facilitate further improvements in the operating system, the laboratory machine includes facilities for manual editing of the dictionary, the phonetics, and the acoustic components of the message, together with visual displays (by means of a cathode ray tube) of each of these aspects. This allowance for editorial intervention is an important developmental feature, but it will of course not be essential to the final reading system. However, during field evaluation, substantial feedback from the separate tests is expected, and what is learned can be structured into on-going modification of the speech-producing program.

FIELD EVALUATION

The purpose of the evaluation tests will be to attempt to answer questions concerning human factors, cost versus benefit, and technical matters (Nye et al., 1972). In pilot studies conducted over the past 18 months we have had the cooperation of faculty and students at the University of Connecticut and at the Veterans Administration Hospital in West Haven, and it is anticipated that in the conduct of the detailed evaluation study we will again rely upon both of these institutions for assistance in test design, in acquisition and scheduling of testees, and in administration of the tests.

Some of the test materials (rhyme tests, for example) are standard and have been used elsewhere for such applications as assessing telephone speech quality. We also intend to use reading comprehension tests similar to those used in the public education system. On the other hand, certain tests will be created specifically for appraising the synthetic speech medium. Various levels of textual material will be presented, in a variety of subject matter, by a range of authors. (Authors' styles and vocabulary are known to produce differing degrees of acceptability among readers.)

Beyond the standardized and specially-designed listening tests, we propose to operate a partial Reading Machine Service for blind students at the University of Connecticut at Storrs, in order to make reasonable estimates of time and expense involved in actual text-to-speech-to-user production, and in order to make a genuine test of the feasibility of the system. The magnitude of demand by students for the synthetic speech recordings of their textbook assignments (using this partially simulated system) should be indicative of the contribution a full-scale installation of a Reading Service will provide (although it is recognized that an innovation such as synthetic speech may encounter initial resistance). Demand itself is, of course, one clear type of acceptability measure.

When enough data has been gathered to permit a comparison of synthetic speech tapes and natural speech recordings, including such factors as demand, production speed, cost, and acceptability, we should arrive at a realistic index of the new system's overall feasibility.

In summary, a bench model of an automated reading system for the blind exists. Pilot tests of the acceptability of the synthetic speech output have been conducted, and the system is now ready for serious evaluation. Cooperative university and veterans' facilities are on hand to contribute their assistance in this enterprise. The evaluation study itself represents a start on moving the system out of the laboratory and into real-life application.

REFERENCES

Cooper, F. S. (1963) Speech from stored data. IEEE, International Convention Record, Pt. 7, 137-149.

Gaitenby, J. H., G. N. Sholes, and G. M. Kuhn. (1972) Word and phrase stress by rule for a reading machine. Record of the 1972 Conference on Speech Communication and Processing. AFCRL Special Reports 131 (New York, IEEE) 27-29.

Mattingly, I. G. (1968) Synthesis by rule of General American English. Ph.D. dissertation, Yale University. (Published as Supplement to Haskins Laboratories Status Report on Speech Research.)

Nye, P. W. (1972) The case for an automated reading service. Record of the 1972 Conference on Speech Communication and Processing. AFCRL Special Reports 131 (New York, IEEE) 16-18.

Nye, P. W., J. D. Hankins, T. Rand, I. G. Mattingly, and F. S. Cooper. (1972) Field evaluation of an automated reading machine for the blind. Record of the 1972 Conference on Speech Communication and Processing. AFCRL Special Reports 131 (New York, IEEE) 23-26.

Physiological Aspects of Certain Laryngeal Features in Stop Production*

H. Hirose,[+] L. Lisker,[++] and A. S. Abramson,[+++]
Haskins Laboratories, New Haven

This study represents one more effort to describe some physiological correlates of certain laryngeal features in stop consonant production. Our specific concern is with what the linguists call the voiced-voiceless, the aspirate-inaspirate, and the implosive-explosive oppositions. For this study two separate experiments were performed: electromyography (EMG) of the laryngeal muscles during production of intervocalic labial stops and fiberoptic observation of the glottis for the same consonants.

The subject was a phonetician who produced labial stops of the five phonetic types represented in the bottom line of Figure 1. From left to right, these are: voiced inaspirates, implosive voiced inaspirates, voiced aspirates, voiceless aspirates, and voiceless inaspirates. Of the various carrier utterance types indicated we shall be talking only of those with the vowel [i] in the second and third syllables.

In the first experiment, conventional bipolar hooked-wire electrodes were inserted into five laryngeal muscles: the interarytenoid (INT), posterior cricoarytenoid (PCA), lateral cricoarytenoid (LCA), cricothyroid (CT), and sternohyoid (SH). The EMG signals were recorded along with acoustic signals and timing marks as the subject read the list of test nonsense words 16 to 20 times each. The EMG signals were then computer-averaged with reference to a line-up point on the time axis representing the beginning of the labial closure interval in each utterance. In this discussion we shall restrict ourselves to events in the immediate neighborhood of the line-up point.

Figure 2 shows an example of the averaged EMG curves for the five laryngeal muscles during production of test words containing the explosive voiced inaspirate [b]. The zero on the abscissa marks the line-up point, and each division repre-

---

*This is a slightly modified version of a paper presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., November 1972.

[+]Also Faculty of Medicine, University of Tokyo.

[++]Also University of Pennsylvania, Philadelphia.

[+++]Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

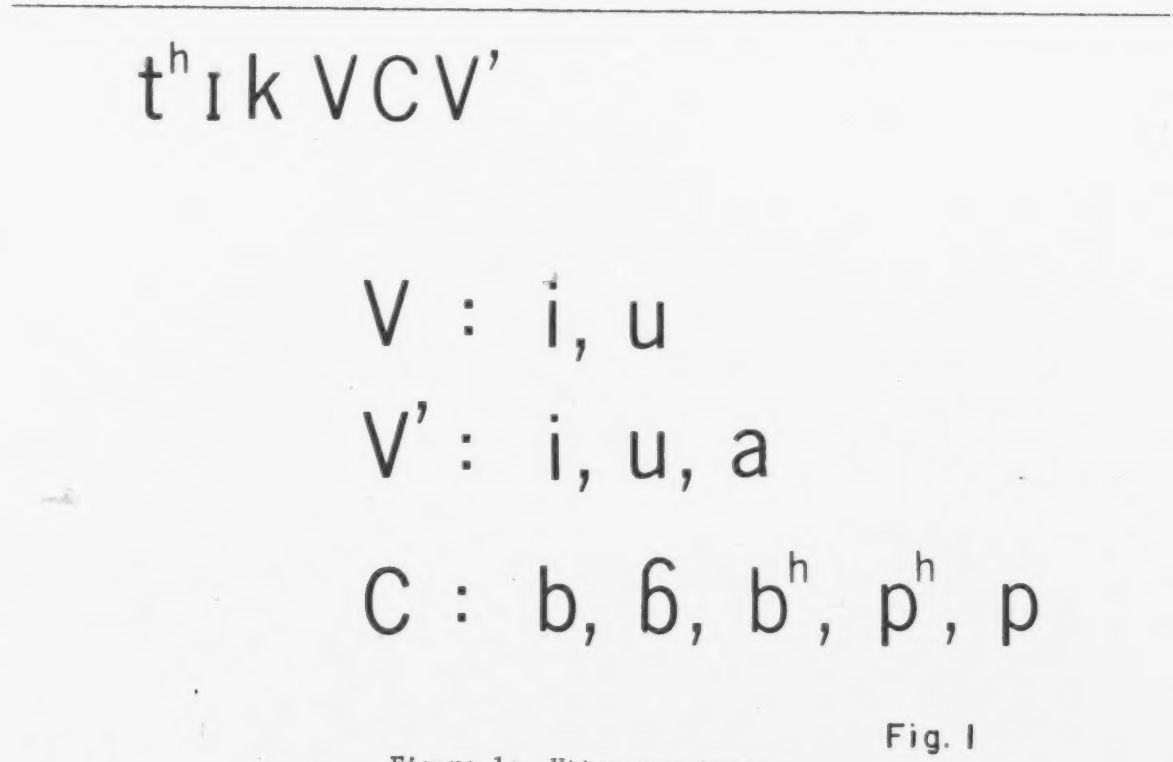


Fig. 1

Figure 1: Utterance types.

sents 100 msec. The EMG curve for the INT shows that this muscle has a steep increase in activity immediately before release of the initial [t^h], which is indicated by a vertical bar at the bottom of the figure. Except for a dip about 200 msec before the line-up point which presumably corresponds to the [k], the INT stays active over the interval containing the labial stop of this phonetic type.

When we then compare the EMG curve for the PCA with the one for the INT, we can see a clear reciprocal relationship between these two muscles. Thus PCA activity remains continuously suppressed after a small peak which is presumably for the [k] in the carrier. The general pattern of LCA activity is more or less similar to that of the INT in this case. The characteristics of the CT and SH will be discussed later.

In Figure 3, the results for the test utterance containing the voiceless aspirates [p^h] are demonstrated. Note, in this case, that there is a marked dip in INT activity starting approximately 100 msec prior to the line-up point and reaching its minimum approximately at the release of the stop. Conversely, marked activity is observed for PCA, starting and peaking reciprocally with the fall in INT activity. In the LCA curve there is a dip in step with INT suppression.

The EMG data in Figures 2 and 3 show marked activity of the abductor muscle in the production of the voiceless aspirates, while activity is suppressed for the voiced inaspirates. The adductor muscle group is, conversely, suppressed

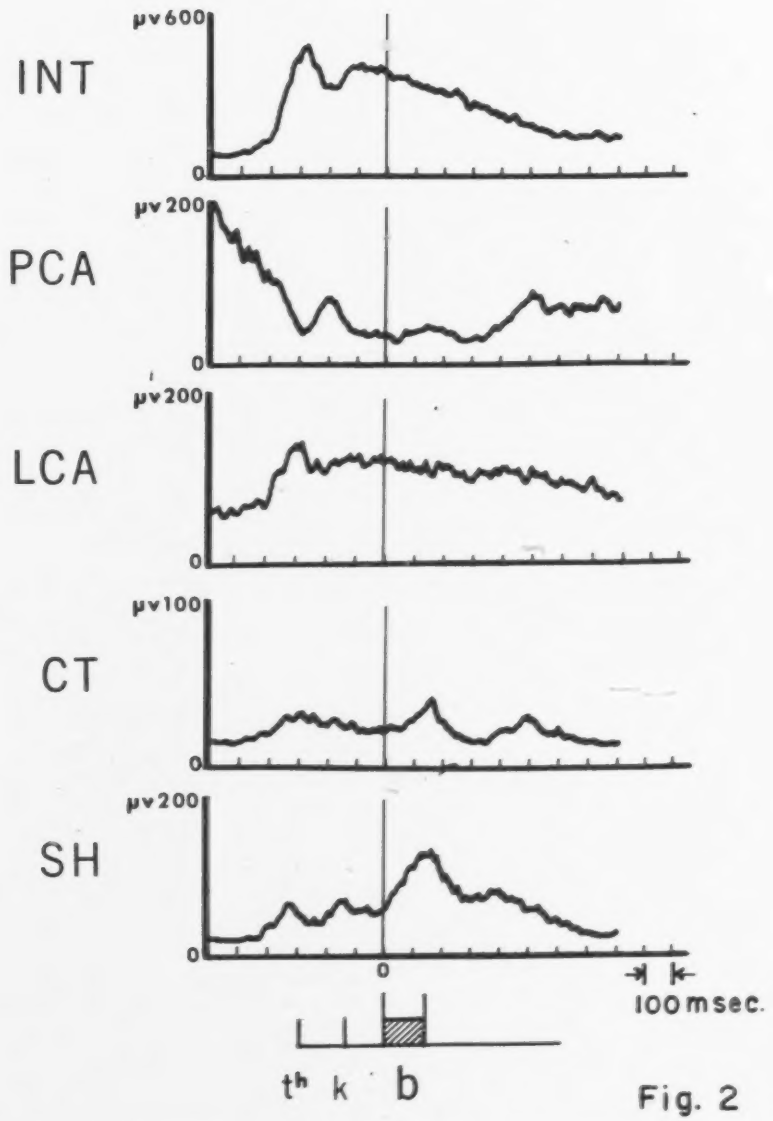


Figure 2: Averaged EMG curves for the five laryngeal muscles for [b]. The time of release of [t^h] and [k] and of the onset and release of [b] are indicated at the bottom of the figure. The shaded interval represents voicing throughout the [b] occlusion.

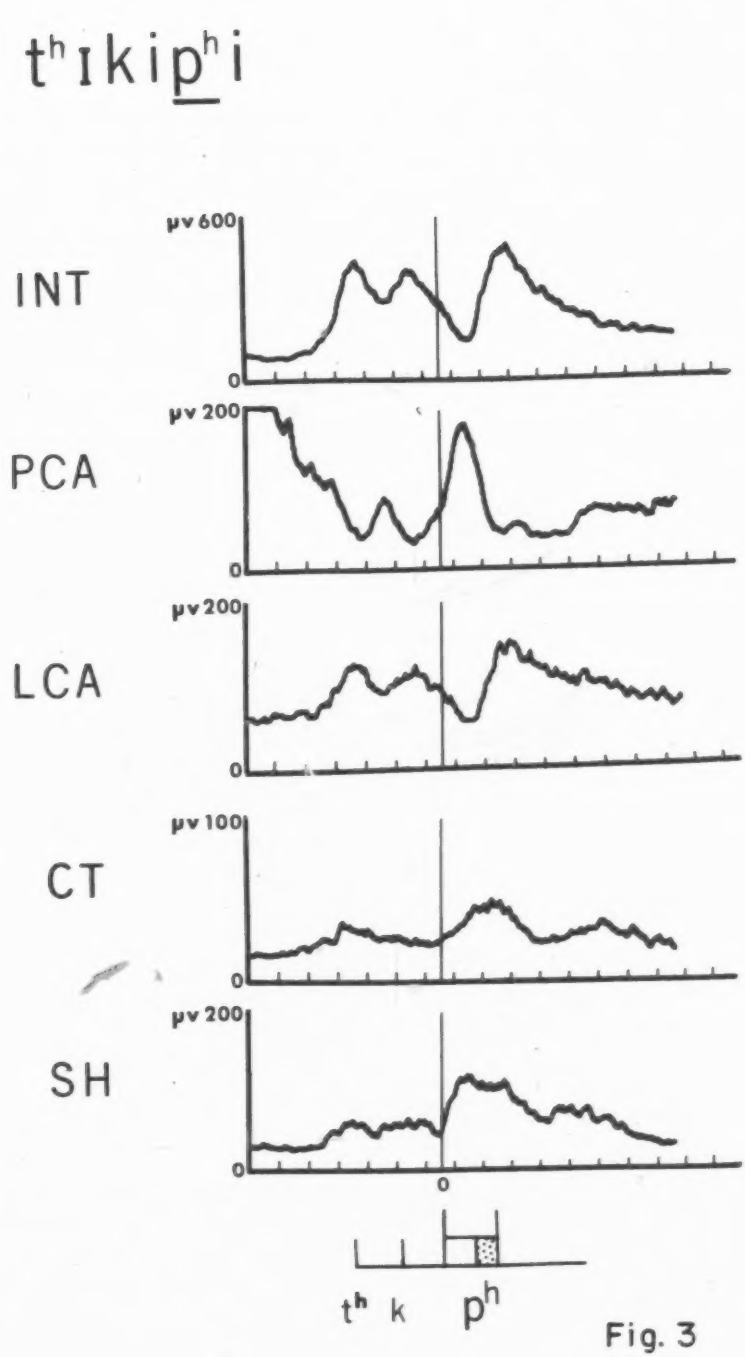


Figure 3: Averaged EMG curves for [p^h]. The times of closure and release and the interval of aspiration for [p^h] are shown.

for voiceless aspirate production but active for the voiced inaspirates.

In Figure 4, averaged EMG curves for three different phonetic types are superimposed for each of the five muscles, allowing us to compare the voiced aspirates, the voiceless aspirates, and the voiceless inaspirates. As we can see, the increased activity of the PCA and the suppression of the INT and LCA, which together can be taken as an indication of open glottis, are most marked for the voiceless aspirates. For the stop types compared in this slide, there is minimal activation of the PCA for the voiceless inaspirates, and its timing is earlier than for the other two classes of stop. There is a similar relation for the reciprocal INT curves.

Comparison of the voiced aspirates and the voiceless aspirates shows a significant difference both in magnitude and timing of the EMG patterns, in that abductor activation and adductor suppression start somewhat earlier and reach higher magnitudes for the voiceless than for the voiced aspirates.

It should be remarked here that, except for the differences relevant to the phonetic differences among the labial stops, the contours of the three curves are similar. This indicates that the pattern of muscle act ity for the carrier portion of the test utterances is quite constant, regardless of the difference in the embedded labial stops. Incidentally, it is also to be noted that both the CT and SH show no differences in EMG contour for the three labial stop types.

Figure 5 compares the voiced inaspirates, the voiced aspirates, and the voiced implosives. Note that there is no appreciable adductor suppression for the voiced inaspirates. The abductor appears to be continuously suppressed in the case of the ordinary voiced stops, but for the implosives a small peak is observed, a finding for which we have yet to discover an explanation.

It is interesting to note that the SH shows a peak for the implosive that occurs earlier than for the other two stop types, one which precedes oral release and presumably is for active larynx lowering. This pattern of SH activity was also found in an earlier experiment involving a native speaker of Sindhi as a subject. In Sindhi, a language of the Indian subcontinent, the implosive feature is distinctive. For this sound the CT also shows a peak that is almost synchronous with the SH peak. This CT activity might be regarded as a tensing of the vocal folds in compensation for a possible pitch drop due to larynx lowering.

In the second part of our study movies of the glottis were taken during production of the same types of test utterances by means of a fiberscope at a film rate of 60 frames per second. Appropriate frame sequences for the labial stops were then examined by the technique of frame-by-frame analysis. The following features were particularly noted:

1) opening and closing movements of the arytenoid
2) interruption and resumption of vocal fold vibration
3) the time course of glottal aperture width as measured at the vocal processes
4) vertical movement of the larynx.

It was observed that all samples of voiced inaspirate sounds, both the ordinary explosive and the implosive, showed no separation of the arytenoids. In the

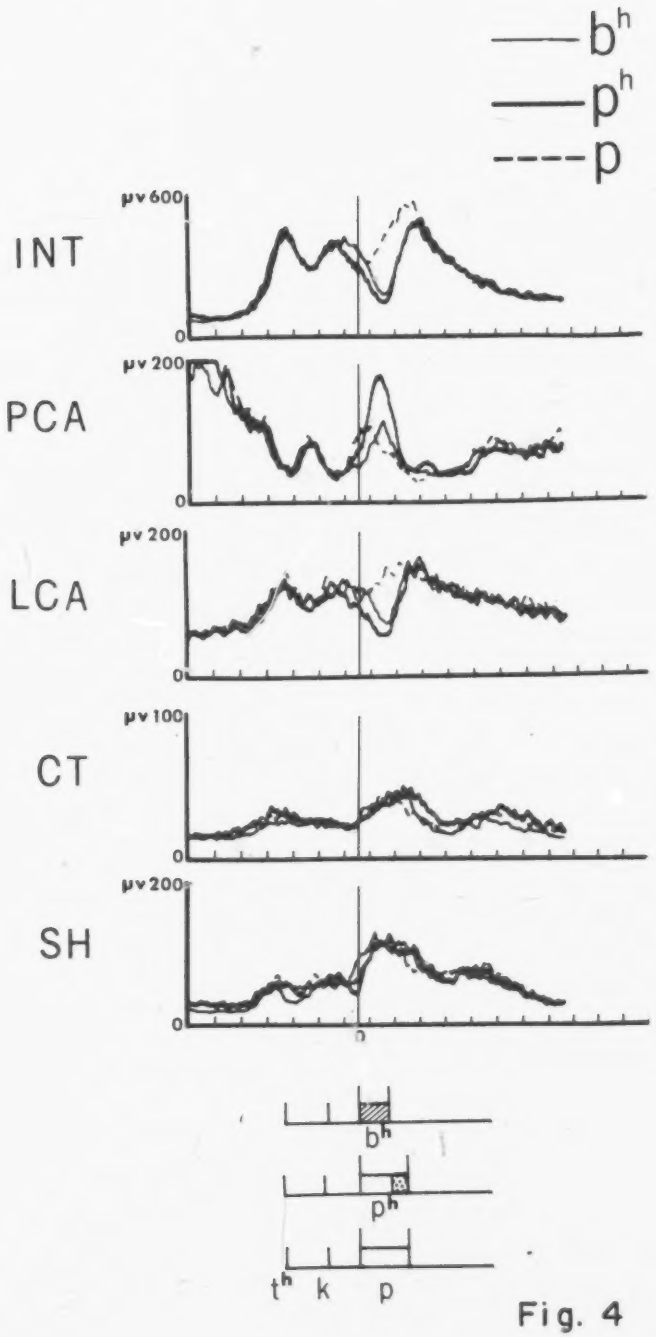


Fig. 4

Figure 4: Comparison of EMG curves for [bʰ], [pʰ], and [p].

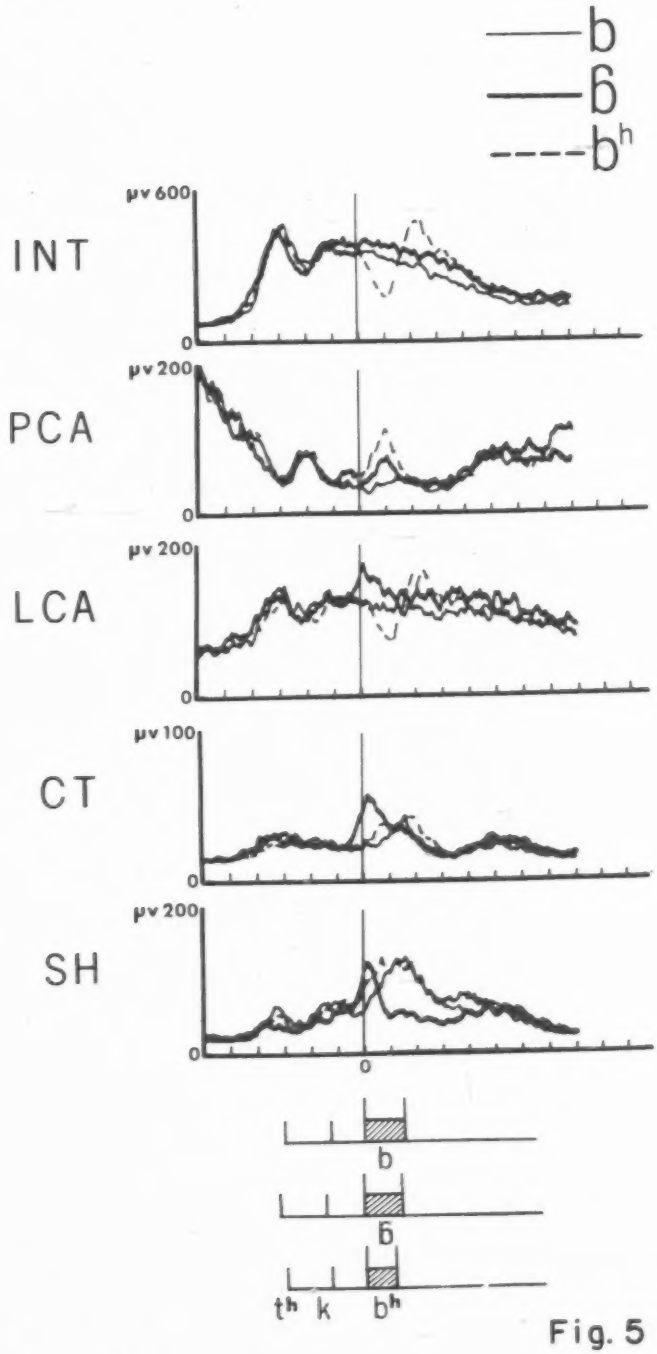


Figure 5: Comparison of EMG curves for [b], [ɓ], and [bʰ].

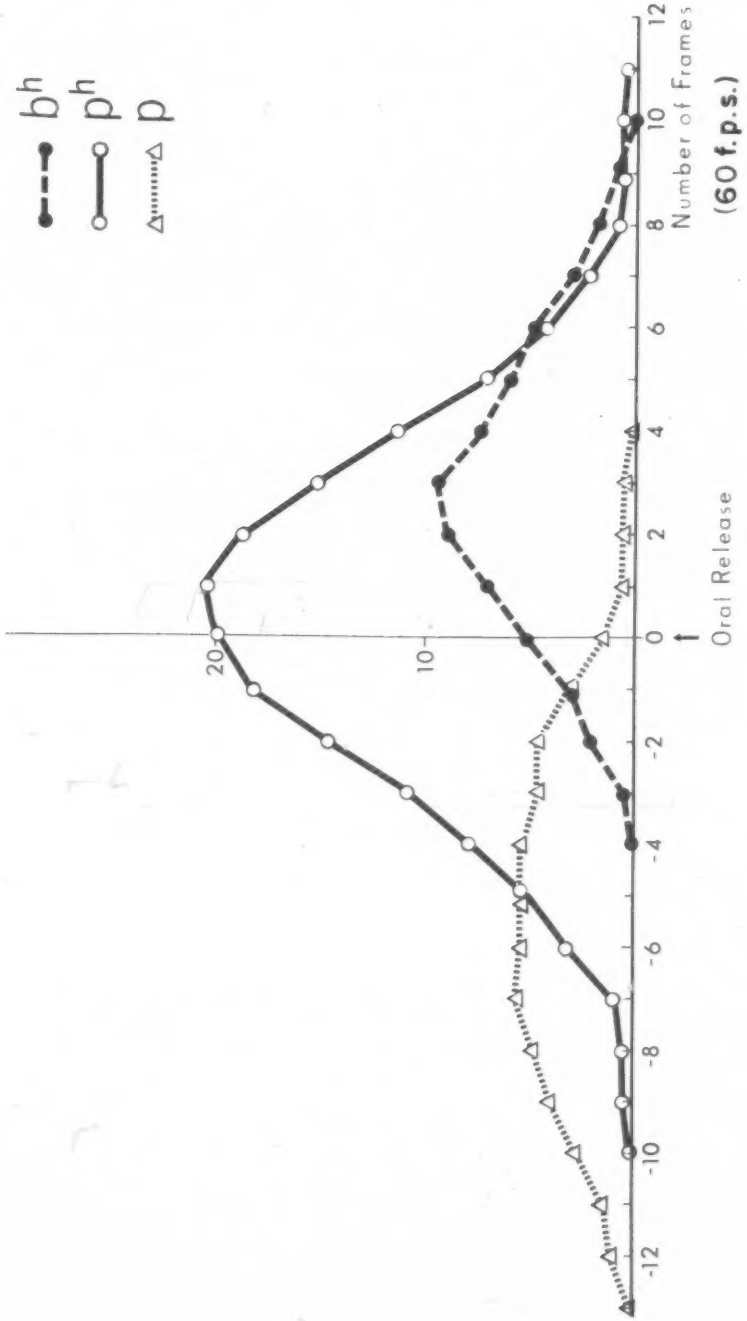


Fig. 6

Figure 6: Time courses of glottal width for [b^h], [p^h], and [p].

case of the implosives the larynx appeared regularly to shift downwards, this being manifested as a rapid tilting of the epiglottis in a postero-inferior direction.

In the phonetic types other than the two voiced inaspirate stops arytenoid separation was always observed. Figure 6 illustrates the average time courses of glottal opening for these three stops relative to oral release. (The unit of glottal width is arbitrary.) The figure clearly shows the difference in temporal course as well as degree of opening of the glottis for the different phonetic types. The glottis appears to open earlier and also to close earlier relative to oral release for the voiceless inaspirates than for the other two, while the maximum glottal width is the smallest for the same stop. For the voiceless aspirates, the glottis starts to open later, reaches its maximum width near oral release, and then gradually closes. The maximum glottal width is largest for this stop. For the voiced aspirates, the glottis appears to open latest, reaching its maximum width somewhat later than for the voiceless aspirates, but closes at almost the same time as for the voiceless aspirates.

If we compare our fiberoptic observation with the EMG data described earlier, there appears to be good agreement between the two sets of results, indicating in particular that the abductor and adductor muscle groups follow coordinated patterns of activity corresponding to the opening and closing of the glottis. The overall conclusion suggested is that the classes of phonetic events we have been examining are produced by rather different laryngeal gestures deriving from different patterns of laryngeal muscle activity. For the voicing and aspiration features these patterns effect differences both in the magnitude of glottal opening and in the timing of glottal adjustment relative to supraglottal events, while for the implosive feature there is lowering of the larynx in synchrony with the oral closure.

Effect of Speaking Rate on Labial Consonant Production: A Combined Electromyographic-High Speed Motion Picture Study

Thomas Gay[+] and Hajime Hirose[++]
Haskins Laboratories, New Haven

ABSTRACT

This experiment used electromyography and direct, high speed motion picture photography, in combination, to describe the effect of speaking rate on the production of labial consonants.

Electromyographic signals from a number of facial muscles were recorded simultaneously with high speed motion pictures of the lips from two subjects. The speech material consisted of syllables containing the consonants /p b m w/ in both CV and VC combinations with the vowels /i a u/.

The major finding of this experiment is that an increase in speaking rate is accompanied by both an increase in the activity level of the muscle and an increase in the speed of movement of the articulators. The data also showed certain manner effects and instances of both subject-to-subject and individual token variability. These findings are discussed in terms of theoretical models of speech production.

It is commonly known that the production of a given phone will vary a great deal depending on the suprasegmental structure in which it is placed. Recent research in this area has been concerned with the question of whether these allophonic variations, in particular those that arise from changes in stress and speaking rate, can be attributed solely to changes in the timing of commands to the articulators.

The earliest model of this type was proposed by Lindblom (1963, 1964). In both spectrographic and cinefluorographic studies, Lindblom found that a destressed vowel, or one produced during faster speech, was accompanied by a change in color toward the neutral schwa. Lindblom's hypothesis was that this neutralization is a consequence of the shorter duration of the vowel, and further, is caused by a temporal overlap of motor commands to the articulators.

[+]Also University of Connecticut Health Center, Farmington.

[++]Also Faculty of Medicine, University of Tokyo.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

In other words, the articulators fail to reach, or undershoot, their targets because the next set of motor commands deflects them to the following target before the first target is reached. Although some later experiments have shown similar undershoot effects for other phones (Gay, 1968; Kent, 1970), a number of other studies have produced results that imply the existence of another mechanism, articulatory reorganization, in the control of, at least, stress. For example, both Harris, Gay, Sholes, and Lieberman (1968) and Harris (1971), in electromyographic studies of stress, found higher muscle activity peaks (greater driving forces) for phones produced in stressed syllables than in non-stressed syllables. A possible consequence of the electromyographic result was later observed by Kent and Netsell (1971) in a cinefluorographic study of tongue movements. Their data suggest that the effect of increased stress is to cause the articulators to move faster, more forcefully, and closer to their intended targets.

Although it is probably safe to conclude that undershoot is at least a component of destressed and faster speech, a general model of speech production based on timing changes alone is too simple. First, it is apparent from earlier experiments that reorganization of the articulatory gesture exists to enable the mechanism to respond actively to, at least some, suprasegmental demands. Second, the concept of undershoot, which was originally proposed to describe vowel articulation, does not lend itself particularly well to the production of consonants, most of which involve movements towards occlusal or constrictive, rather than spatial, targets. The experiment reported here was concerned with the following questions: does articulatory reorganization extend to variations that arise from changes in speaking rate, and can a mechanical model of the kind proposed by Lindblom (1963) apply to the production of both vowels and consonants? The specific purpose of this experiment was to determine the effects of speaking rate on the production of labial consonants spoken in various vowel environments. The experimental approach utilized the combined techniques of electromyography and direct high speed photography to obtain information about both the forces that cause the articulators to move and the movements that result, simultaneously, on the same utterance.

METHOD

Subjects and Speech Material

The subjects were two adult males, both native speakers of American English. The speech material for one subject (DL) consisted of the labial consonants /p b m w/ in both CV and VC (except /w/ which was in only CV) combination with the vowels /i a u/. Each of the syllables was placed in a word (e.g., keeper, appeal), which, in turn, was placed in a sentence. The master list contained 21 different words. For the second subject (TG), a more symmetrical frame was used. Each of the consonants was placed in either [kVCə] or [kəCV] (again, except for /w/) preceded by the carrier, "It's a...." Also, since the data analyzed for the first subject did not show any interesting or consistent manner differences for /m/, this consonant did not appear in the second list. For both subjects, the words were random ordered into four different lists. The lists were repeated four times, in sequence, for a total of sixteen repetitions at each of two different speaking rates. The speaking rates were either moderate or fast and were controlled by training the subject to speak at what he considered comfortable rates. The subject's performance was monitored continuously throughout the run.

Electromyography

For both subjects, conventional hooked wire electrodes were inserted into muscles that control both lip and jaw movements. These muscles are listed in Table 1. Although all muscle locations showed adequate firing levels at the time of electrode insertion, some muscle locations deteriorated, at one time or another, during the run. The extent to which this occurred is also indicated in Table 1.[1]

TABLE 1: EMG electrode locations.

Subject DL	Subject TG
Orbicularis Oris - superior, medial (OOSM)	Orbicularis Oris - superior, medial (OOSM)
Orbicularis Oris - superior, lateral (OOSL)*	Orbicularis Oris - inferior (OOI)
Orbicularis Oris - inferior (OOI)	Quadratus Labii Inferiorus (QLI)
Levator Anguli Oris (LAO)	Mentalis (MEN)
Buccinator (BUC)	Anterior Belly Digastric (AD)*
Depressor Anguli Oris (DAO)	
Internal Pterygoid (IP)*	

*Analyzed for motion picture segment only.

The basic procedures were to collect EMG data for a number of tokens of each utterance and, using a digital computer, to average the integrated EMG signals at each electrode position. The EMG data were recorded on a multichannel instrumentation tape recorder together with the acoustic signal and a sequence of digital code pulses (octal format). These pulses were used to identify each utterance for the computer during processing. A more detailed description of both the data recording and data processing techniques can be found elsewhere (Hirose, 1971; Port, 1971).

Direct High Speed Photography

High speed motion pictures of lip and jaw movements were recorded with a 16 mm Milliken camera, set up to run at 128 fps. Both full-face and lateral views of the lips and jaw were recorded by placing a mirror, set at a 45 degree angle, beside the subject's face. The motion picture and EMG data were synchronized by an annotation system constructed for the purpose that displayed

[1] It is interesting to note further, that for Subject TG, the internal pterygoid muscle showed only resting potentials for speech, even though correct electrode placement was ascertained for other functions (clenching, for example).

the octal code pulses on an LED device placed in the path of the camera. This display was also driven by a signal from the camera to count individual frames between the octal codes. A diagram of the EMG and motion picture instrumentation is shown in Figure 1.

The combined EMG and high speed motion picture data were recorded at the beginning of the run, after which the EMG part of the experiment continued. Prior to the beginning of the run, white reference dots were painted on the nose and lower edge of the subject's jaw. A ruler was also fixed to the mirror so that the lip and jaw movements could be converted to actual distances.

The films were analyzed by frame-by-frame measurements of vertical jaw opening and vertical lip opening at the midline.[2] The EMG data from the motion picture part of the run were processed separately from the remainder of the run and then compared with those data to see if the individual tokens were typical of the average. Our criterion for acceptance of an individual token was that its peak did not exceed the maximum or fall below the minimum of the averaged tokens.

RESULTS

Electromyography

Results of the electromyographic analyses are summarized in Tables 2 and 3. These tables show the peak muscle activity levels for each muscle and utterance for both subjects and both speaking rates.

The values for the OOSM, GOSL, OOI, MEN, and IP represent the peak heights of muscle activity levels for the closing segment of the gesture while the values for the AD, BUC, DAO, LAO, and QLI represent the peak heights for the opening segment.

For both subjects, the speech produced during the faster speaking rate condition was, on the average, one-third shorter in duration than the speech produced during the normal condition. These differences are based on measurements made from the complete sentence, i.e., test syllable plus carrier.

One of our major concerns in combining EMG with high speed motion picture photography was the question of whether the EMG curve for the single token motion picture run was a typical one, in other words, one compatible with the average. As the values in Tables 2 and 3 indicate, this was almost without exception the case. For those muscles characterized by strong activity levels, the single token values followed the averaged values, in both direction and magnitude. Almost all of the other muscle locations showed the same patterns; however, since the peak values of these muscles were somewhat lower, the comparisons are not equally valid.[3]

---

[2] Lip spreading measurements (horizontal distance between the corners) were also made for /w/.

[3] For subject DL, two instances (/im, wi/ - OOI), and for subject TG, one instance (/pu/ - MEN), occurred where the single token values did not follow the averaged values.

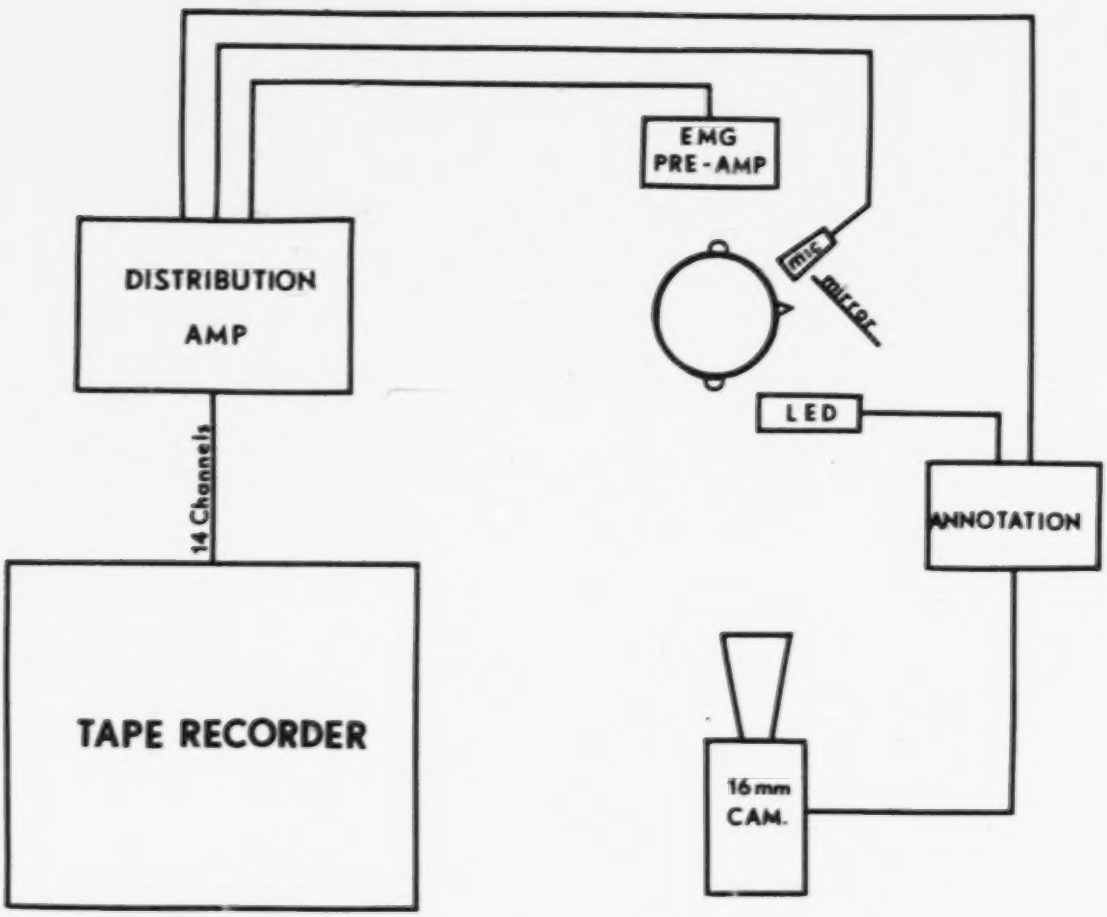


Figure 1: Block diagram of EMG and high speed motion picture recording system.

TABLE 2: Averaged and single token (in parentheses) peak EMG values for subject DL. Values for the moderate speaking rate are in the left column and values for the fast speaking rate are in the right column of each cell.

	OOSM	OOSL	OOI	LAO	BUC	DAO	IP
pi	83-127 (60-190)	(25-52)	58-69 (110-40)	27-38 (38-52)	46-126 (354-550)	690-710 (450-975)	(130-550)
pa	56-75 (60-83)	(25-125)	77-77 (50-65)	27-27 (33-60)	76-304 (30-142)	451-564 (605-950)	(150-690)
pu	67-71 (150-170)	(95-195)	70-74 (65-135)	26-37 (20-35)	77-191 (50-350)	504-616 (820-960)	(125-700)
ip	54-82 (75-155)	(65-75)	37-47 (40-50)	22-26 (60-80)	65-156 (85-175)	626-788 (710-920)	(75-650)
ap	53-67 (65-105)	(60-95)	34-41 (35-40)	35-39 (35-40)	112-183 (65-410)	915-1009 (960-980)	(65-620)
up	77-88 (95-115)	(80-125)	95-111 (105-180)	138-149 (145-185)	135-242 (75-85)	837-1167 (850-960)	(130-600)
bi	78-114 (70-180)	(65-90)	59-69 (65-130)	27-68 (30-40)	81-230 (120-180)	671-965 (920-1020)	(75-700)
ba	99-108 (60-155)	(40-80)	63-69 (75-105)	39-61 (50-60)	132-514 (75-650)	730-854 (900-950)	(125-695)
ib	65-85 (40-160)	(40-45)	48-62 (40-60)	191-124 (60-240)	58-114 (60-350)	532-717 (225-950)	(125-650)
ab	64-80 (40-160)	(70-80)	47-54 (40-55)	81-47 (80-70)	130-134 (95-120)	763-938 (740-890)	(130-595)
mi	110-129 (45-190)	(35-95)	66-72 (42-65)	59-67 (55-85)	73-290 (40-560)	698-836 (475-980)	(160-480)
ma	85-113 (95-195)	(65-130)	53-58 (60-95)	37-65 (85-105)	179-560 (65-705)	662-704 (575-1145)	(140-710)
im	63-70 (70-110)	(35-65)	57-54 (45-55)	31-35 (30-35)	63-185 (50-170)	654-630 (550-950)	(130-620)
am	74-76 (55-170)	(80-95)	67-69 (70-80)	30-32 (50-55)	70-108 (190-275)	809-918 (640-950)	(180-390)
wi	68-78 (60-130)	(55-105)	88-69 (75-155)	22-26 (25-200)	124-145 (100-150)	318-367 (75-775)	(105-705)
wa	58-60 (55-60)	(65-150)	33-37 (55-95)	25-34 (25-30)	113-184 (105-200)	213-505 (50-650)	(125-485)

TABLE 3: Averaged and single token (in parentheses) values for subject TC. Values for the moderate speaking rate are in the left column and values for the fast speaking rate are in the right column of each cell.

	OOSM	OOI	QLI	MEN	AD
pi	381-555 (414-451)	82-111 (91-95)	36-37 (43-46)	34-26 (47-51)	(20-22)
pa	323-634 (425-502)	62-101 (91-161)	33-36 (32-43)	37-39 (39-202)	(68-90)
pu	369-588 (409-523)	98-113 (83-113)	29-31 (32-40)	36-26 (42-52)	(24-41)
ip	528-580 (431-540)	90-111 (141-151)	40-49 (29-38)	61-65 (65-70)	(57-112)
ap	542-583 (551-563)	92-115 (111-112)	76-80 (81-108)	64-67 (81-89)	(35-45)
up	346-435 (305-420)	110-186 (91-133)	32-39 (35-40)	19-19 (39-40)	(65-82)
bi	357-532 (305-529)	56-94 (63-97)	25-27 (27-43)	29-22 (44-45)	(10-13)
ba	299-458 (458-478)	55-126 (41-80)	23-26 (24-27)	41-30 (47-47)	(57-101)
bu	317-456 (283-409)	84-106 (72-127)	20-22 (21-24)	36-20 (52-59)	(10-46)
ib	432-508 (403-409)	52-84 (61-66)	26-29 (29-31)	55-57 (42-65)	(32-52)
ab	535-588 (409-431)	61-75 (55-62)	42-54 (39-48)	62-71 (71-77)	(10-38)
ub	307-347 (376-447)	94-188 (94-233)	16-21 (27-35)	23-29 (31-39)	(11-41)
wi	244-440 (245-321)	80-143 (80-105)	15-20 (19-27)	7-8 (7-10)	(10-32)
wa	277-441 (218-300)	88-185 (80-147)	14-32 (13-21)	7-11 (7-10)	(32-41)
wu	277-392 (218-343)	89-149 (91-169)	12-18 (18-27)	10-8 (7-28)	(10-46)

For both subjects, the major effect of an increase in speaking rate was an increase in the activity level of the muscle. Generally speaking, this increase occurred for all muscles and all utterances, in other words, for the immediate consonant gesture itself, as well as for the lip opening that extends through the adjacent vowel.[4] For those muscles characterized by strong activity levels, these increases were on the order of anywhere from 25 percent to 100 percent.

These differences in muscle activity levels indicate that the control of speaking rate requires more than just a simple adjustment of the timing of motor commands. Rather, it appears that the labial consonant gesture is also reorganized at the muscle command level. Although changes in timing are present, the primary physiological correlate of increased speaking rate seems to be a generalized increase in articulatory effort.

Whereas the speaking rate effects were consistent for both subjects, other effects, both contextual and suprasegmental, were quite variable. For example, examination of the orbicularis oris data for subject DL shows that, for the most part, muscle activity levels were higher for /p b m/ produced before a stressed vowel than for /p b m/ produced after a stressed vowel. (The other muscles do not show any consistent stress effects.) For subject TG, on the other hand, stress contrasts for /p b/ in combination with the vowels /i a/ show exactly opposite effects: consistently higher EMG activity (OOS, OOI, QLI) for the poststressed position; /u/ showed small, probably inconsequential effects the other way (prestressed). For both subjects, these effects occurred across changes in speaking rate. Likewise, subject TG showed higher OOS, OOI, QLI activity levels for /p/ as opposed to /b/, while no consistent differences were evident for subject DL. This latter variability has been shown in a number of earlier studies (Harris, Lysaught, and Schvey, 1965; Fromkin, 1966; Tatham and Morton, 1968).

One other interesting finding is worth mentioning. Although the internal pterygoid location for subject DL was not usable for the entire EMG run, it was stable for the motion picture segment. The fact of the single token notwithstanding, the effect of speaking rate on the activity of this muscle was dramatic. Activity levels for all utterances increased from what might be considered resting levels for normal speaking rate to very high peaks for fast speaking rate. This is especially interesting in light of the fact that subject TG did not seem to use this muscle at all for speech.

To summarize at this point then, the major effect of an increase in speaking rate on labial consonant production is a generalized increase in the activity levels of the muscles; this in turn indicates an overall increase in articulatory effort for these consonants during faster speech.

Lip Movements

Figure 2 shows typical lip movement curves for /p b w/ of subject DL. For /p b/, these graphs show that the rates of lip opening and closing are faster for the faster speaking rate condition, while lip closure duration remains

---

[4]The only exceptions were /ab/ - LAO for subject DL and /bi, ba/ - MEN for subject TG.

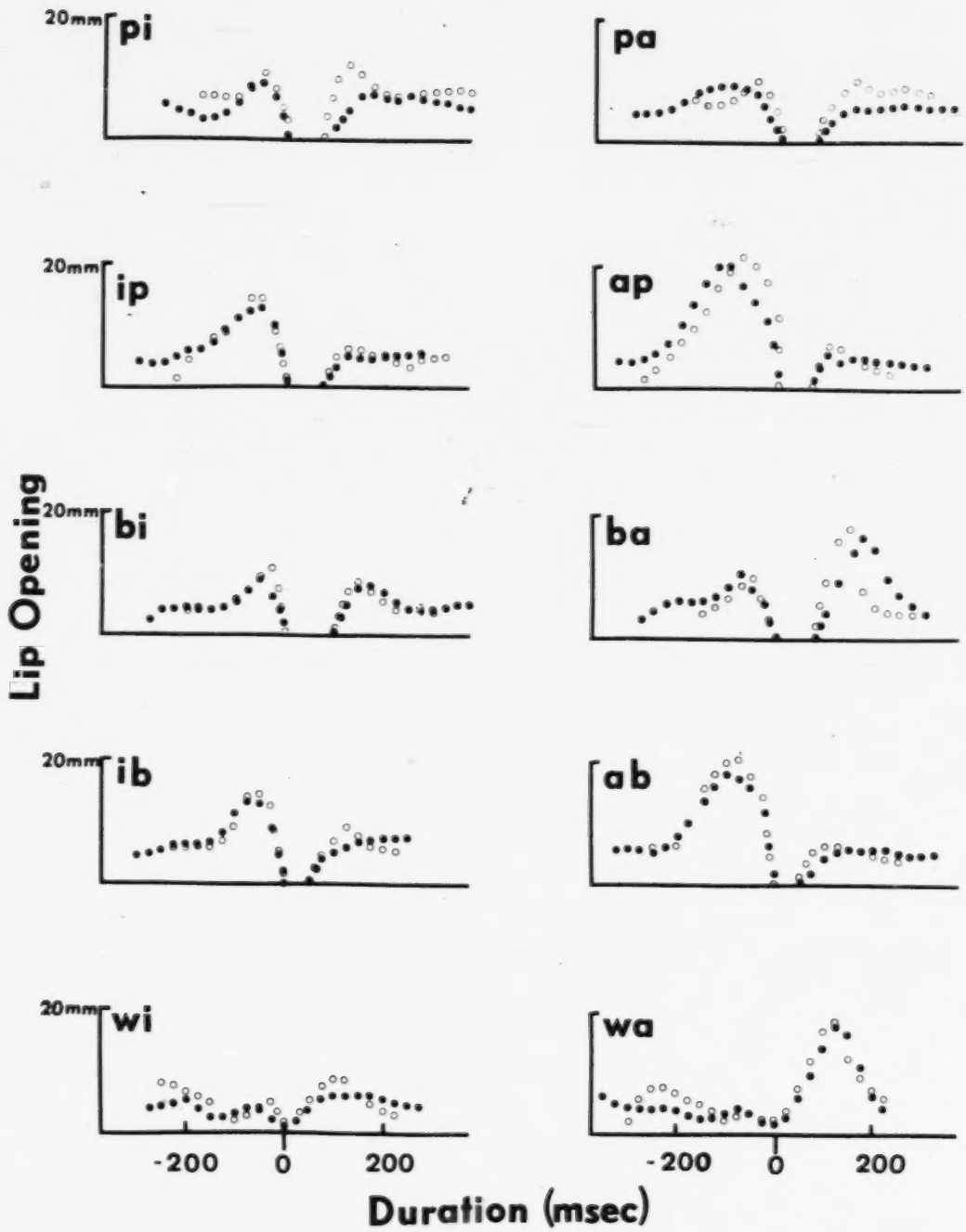


Figure 2: Vertical lip opening curves for subject DL. "0" on the abcissa represents the point of lip closing for /p b/ and the point of minimum opening for /w/. The moderate rate is represented by filled circles and the fast rate by unfilled circles. Data points are plotted at 25 msec intervals except from ± 25 msec of lip closure, where they are plotted at 10 msec intervals.

essentially the same across both rates. This was generally the case for all utterances, although in two instances (both involving /u/), rates of movement were similar for both conditions. However, in no case were the rates of lip movement ever slower for the faster speaking rate condition. The data for subject TG were essentially the same, with only one instance (/pu/) showing similar rates.

These differences in lip movements are consistent with the EMG data and show that in order for the gesture to be completed during faster speech, it must be done faster, and with greater articulatory effort. This increase in effort for the consonant gesture carries over to the adjacent vowel as overshoot in lip opening. This greater amount of lip opening during faster speech occurs primarily for the stressed vowel, regardless of whether it precedes or follows the consonant. Although overshoot was present for some of the unstressed vowels, it was not a particularly strong or consistent feature.

Figure 2 also shows typical lip movement curves for /w/. Here the targets for lip opening and closing are essentially the same across changes in speaking rate. This indicates, again, that an increase in speaking rate causes the articulators to move faster and more forcefully toward these targets. Although this finding for the stops is not unexpected, the data for /w/ are somewhat surprising. Whereas stop consonant production involves an occlusal target, one that must be reached in order to produce the sound, /w/ involves only a spatial target, and, theoretically can be undershot. One possible explanation for the lack of /w/ undershoot, though, is that /w/ might be characterized by an acoustic steady state, and thus, would require an invariant target position.

As might be predicted from the EMG data, the lip movement curves did not show any consistent stress or contextual effects. Any effects that might have been evident for the averaged EMG data did not show any corresponding differences in lip opening or closing. These subtle effects, if indeed they are even reflected by changes in the pattern of lip movements, are probably masked by the variability inherent in single token analyses.

Based upon both the EMG and motion picture data, it would appear that Lindblom's (1963) undershoot model cannot be applied to the production of labial consonants. Although changes in timing are present, the primary physiological correlate of increased speaking rate is an increase in effort and consequently, a faster articulatory movement. As was mentioned before, this is not necessarily an unexpected result for the stops, since these phones require an occlusal rather than a spatial target, and thus, cannot in a strict sense be undershot (except, of course, in terms of decreased closure duration, which also does not occur). The data for /w/, however, are unexpected since /w/ does involve a spatial target.

Jaw Movement

Although the EMG levels for the muscles that control jaw opening and closing (anterior belly of the digastric and internal pterygoid) showed some increase for the faster speaking rate condition, the jaw movement data did not show any clear speaking rate effects. There were no consistent differences in either rate or degree of jaw opening or closing, i.e., there were no consistent undershoot effects for the consonant, or overshoot effects for the vowel, as a function of speaking rate. Although these inconsistencies might be due to the variability inherent in single token measurements, or for that matter, to the coordinate system itself (movement inferred from superficial measurements), the

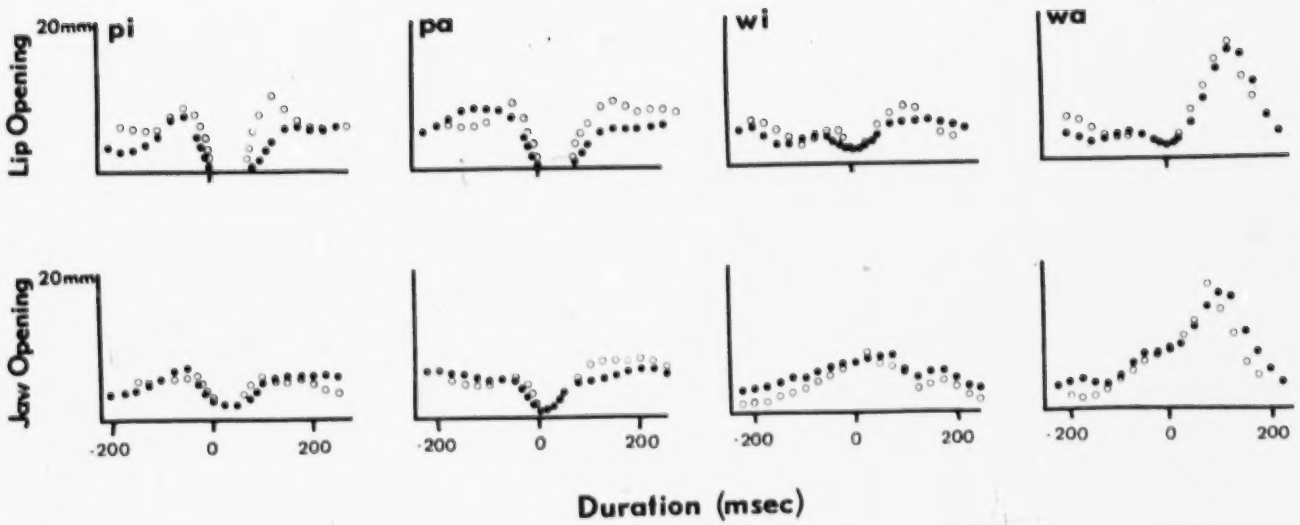


Figure 3: Vertical lip and jaw opening curves for subject DL. "0" on the abcissa represents the point of lip closing for /p/ and the point of minimum opening for /w/. The moderate rate is represented by filled circles and the fast rate by unfilled circles. Data points are plotted at 25 msec intervals except from ± 25 msec of lip closure, where they are plotted at 10 msec intervals.

most likely explanation is that the jaw, unlike the lips, does not need to reach a specific target during the production of a labial consonant, and thus, can be more susceptible to mechanical or inertial factors.

Although the jaw movement curves did not show consistent speaking rate effects, they did show interesting contextual effects. Figure 3 shows lip movement data replotted against jaw movement curves for words containing /p/ and /w/. This figure shows, that for /p/ jaw movement is more or less locked to lip movement, i.e., when one is closing so is the other (this was also the case for /b m/). Lip and jaw coordination for /w/, however, behaved quite differently--jaw movement was much more independent of lip movement, anticipating the following vowel by opening for it during lip closure for /w/. This phenomenon was evident for both subjects, and in each case, the starting point for jaw opening preceded the point of maximum lip constriction by approximately 200 msec.

DISCUSSION

The major finding of this experiment is that for labial consonant production, an increase in speaking rate is accompanied by both an increase in the activity level of the muscles and an increase in the speed of movement of the articulators. Both of these effects are consequences of an increase in articulatory effort. Although these results easily fit into a target-based view of speech production, they do not fit at all into a simple physiological model of the suprasegmental structure of speech.

Lindblom's (1963) original undershoot hypothesis was proposed to account for changes in both stress and speaking rate; that is, his model predicts that undershoot would occur for both destressed and faster speech. Indeed, this seems to be the case for vowels. Both Lindblom (1963) and Kent and Netsell (1971) found that the effect of increased stress is to cause the articulators, specifically the tongue, to move closer to its intended target. Lindblom (1964) also showed the same effect for slower speaking rates, as did Kent and Moll (1972), whose data suggest the same trend for lingual consonants as well as for vowels.

The data of our experiment, however, show that the production of labial consonants is not controlled in the same way as vowels and, perhaps, lingual consonants. For labial consonants, an increase in speaking rate is not accompanied by undershoot, or any corollary change in lip closure duration; rather, the articulatory movement is reorganized at the motor command level in much the same way it is for increased stress, i.e., in the form of greater articulatory effort. Not only does this suggest the existence of more than one mechanism employed in the control of speaking rate but, moreover, that stress and speaking rate variations are not simply covarying components of the same overall structure. Instead they appear to be two features which are controlled by two separate mechanisms.

The data of this experiment show instances of both subject-to-subject and individual token variability. The most interesting subject differences had to do with the EMG measures of the stress and voicing contrasts. These differences were more than likely real since the muscle activity patterns were consistent for utterance versus subject contrasts. The extent of this type of variability, though, is perhaps best illustrated by the data for the internal pterygoid

muscle. For subject DL, this muscle showed rather large speaking rate effects, while for subject TG, the internal pterygoid was not even used for speech. These variations would seem to indicate, among other things, that physiological data of this type should be handled on an individual, nonpooled basis.

The other type of variability apparent in our data was that for jaw opening and closing. As was mentioned earlier, these inconsistencies are probably due to the compounding effects of single token analysis and the fact that the jaw is under less severe constraints than the lips during labial consonant production.

The data of this experiment preclude the hypothesis that the suprasegmental feature of speaking rate is controlled solely by changes in the timing function of the motor commands. It is apparent that an additional, active mechanism is employed in the production of, at least, the labial consonants. However, the extent to which this mechanism operates and the question of whether it operates by feature or by phoneme, cannot be answered without additional data on the way in which the movements of the peripheral mechanism are coordinated with those of the tongue and jaw.

REFERENCES

Fromkin, V. A. (1966) Neuromuscular specifications of linguistic units. Lang. Speech 9, 170-199.

Gay, T. (1968) Effect of speaking rate on diphthong formant movements. J. Acoust. Soc. Amer. 44, 1570-1573.

Harris, K. S. (1971) Vowel stress and articulatory reorganization. Haskins Laboratories Status Report on Speech Research SR-28, 167-177.

Harris, K. S., G. F. Lysaught, and M. M. Schvey. (1965) Some aspects of the production of oral and nasal labial stops. Lang. Speech 8, 135-137.

Harris, K. S., T. Gay, G. N. Sholes, and P. Lieberman. (1968) Some stress effects on the electromyographic measures of consonant articulations. In Reports of the 6th International Congress of Acoustics, ed. by Y. Kohasi. (Tokyo: International Council of Scientific Unions) 1-9.

Hirose, H. (1971) Electromyography of the articulatory muscles: Current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-85.

Kent, R. D. (1970) A cinefluorographic-spectrographic investigation of the component gestures in lingual articulation. Ph.D. thesis, University of Iowa.

Kent, R. D. and R. Netsell. (1971) Effects of stress contrasts on certain articulatory parameters. Phonetica 24, 23-44.

Kent, R. D. and K. L. Moll. (1972) Cinefluorographic analysis of selected lingual consonants. J. Speech Hear. Res. 15, 453-473.

Lindblom, B. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773-1781.

Lindblom, B. (1964) Articulatory activity in vowels. Speech Transmission Laboratory Quarterly Progress and Status Report (Stockholm: Royal Institute of Technology) STL-QPSR 2, 1-5.

Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

Tatham, M. and K. Morton. (1968) Some electromyography data towards a model of speech production. Occasional Papers, Language Center, University of Essex 1, 1-59.

Stop Consonant Voicing and Pharyngeal Cavity Size*

Fredericka Bell-Berti[+] and Hajime Hirose[++]
Haskins Laboratories, New Haven

Aerodynamic forces require an increase in vocal tract volume during stop consonant occlusion if glottal pulsing is to proceed during the period of consonant closure. Previous research has shown that such volume increases do occur for voiced plosives in medial position in American English. Two modes have been postulated for producing this increase. We shall call the first of these passive pharyngeal enlargement, in which decreased muscle activity in certain muscles implies an increase in pharyngeal cavity size for voiced stop consonant production. We shall call the second mode active pharyngeal enlargement, in which increased muscle activity, in muscles which are not involved in passive enlargement, implies increased pharyngeal cavity size for voiced stop production.

Figure 1 is a schematic representation of the two mechanisms for pharyngeal cavity enlargement. The figure to the left represents a cross-section of the pharynx, while that to the right represents a mid-sagittal section of the pharynx. The arrows marked "P" indicate possible dimensions for passive enlargement, while the arrows marked "A" indicate possible dimensions for active enlargement. The solid arrow represents an enlargement dimension whose mechanism has been specified (Bell-Berti and Hirose, 1972). The dotted arrows represent dimensions of pharyngeal expansion whose mechanisms have been suggested (Perkell, 1969; Kent and Moll, 1969), but which have not been finally specified.

Lateral and posterior movement of the pharyngeal walls might be accomplished passively, with lower activity in the superior and middle pharyngeal constrictors and the anterior and posterior faucal pillar muscles for voiced stop than for voiceless stop consonant production. Depression of the larynx and antero-inferior movement of the base of the tongue might be achieved actively, by increased activity of the infrahyoid musculature, particularly the sternohyoid muscle, for voiced, as opposed to voiceless, stop consonant

---

*Paper presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., November 1972.

[+]Also the Graduate School of the City University of New York and Montclair State College, Upper Montclair, N. J.

[++]Also Faculty of Medicine, University of Tokyo.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

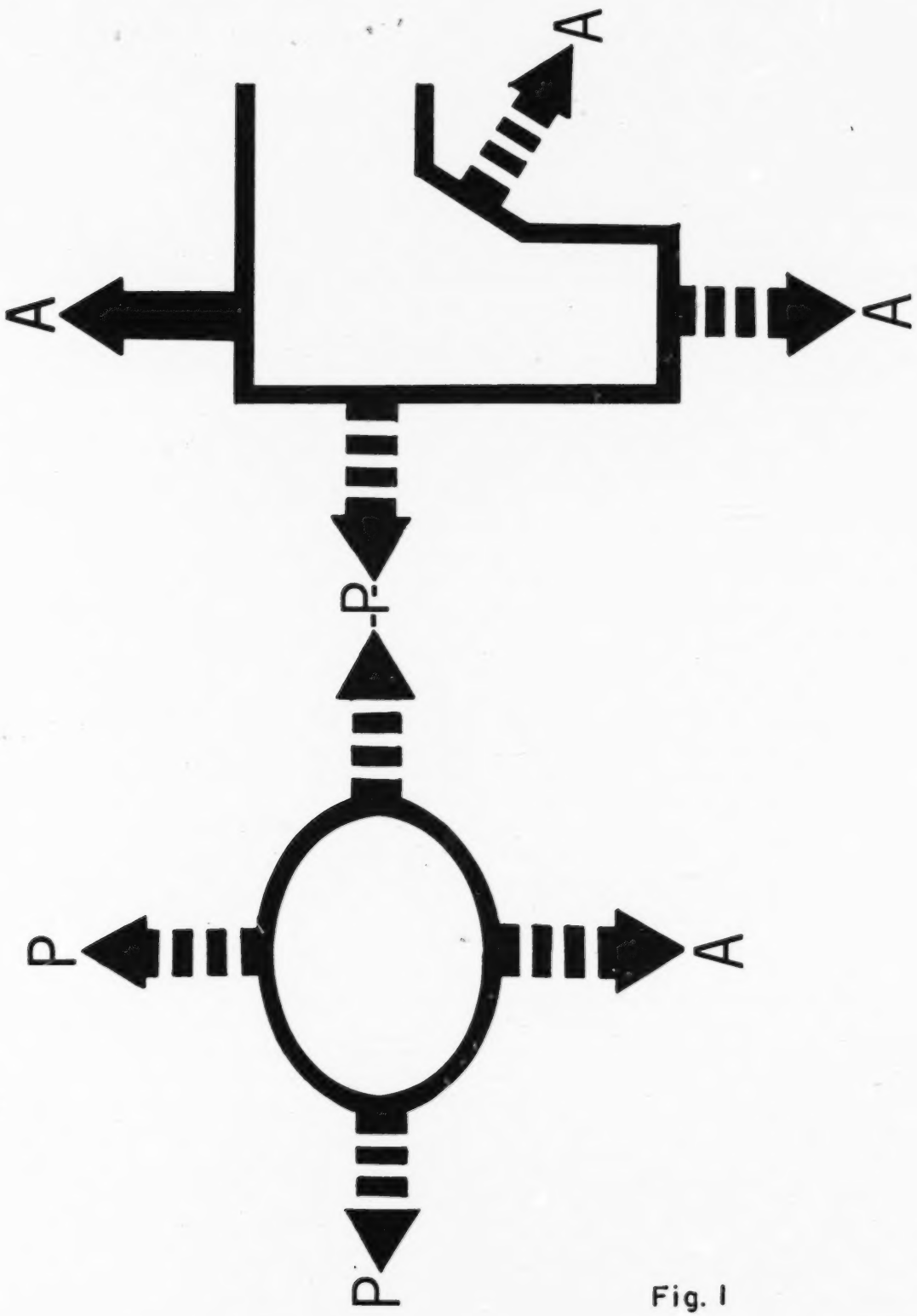


Fig. I

production. The remaining dimension is that of increased velar elevation for voiced stops, which we have previously shown to be accomplished by increased activity of the levator palatini muscle (Bell-Berti and Hirose, 1972).

We obtained simultaneous electromyographic recordings from these six muscles for three speakers of American English. Our stimuli were nonsense disyllables which contrasted the three stop cognate pairs of English in different vowel environments. The nasal consonants /m/ and /ŋ/, which either preceded or followed the stops, were included for another part of the study. Twenty-seven utterance pairs were produced (all beginning with /f/ and ending with /p/, for example: /fapmap-fabmap/, /fimkip-fimgip/ and /fuŋtup-fuŋdup/. The EMG signals were rectified, integrated, and computer-averaged, using the system described by Port (1971). The line-up point for averaging was taken as the boundary between the stop and nasal consonants, determined from an oscillographic record of the audio signal. Spectrographic inspection of the acoustic signal for each subject indicated that voicing proceeded through the period of stop closure for the voiced stops in our samples.

RESULTS

Peaks of EMG activity were found for all of the stop consonants for the levator palatini muscle. We inspected the EMG activity of each muscle at the time of peak levator palatini activity and compared the level of activity associated with stop articulation for each of our 27 stimulus pairs. We assigned the muscles to our passive or active pharyngeal expansion modes and then tested to see if our hypotheses were accepted. When the difference in EMG potential supported the hypothesis for that muscle, a value of "1" was assigned to the muscle for that contrasting pair. When the difference failed to support the hypothesis a value of "0" was assigned and when there was no difference, a value of "1/2" was assigned.

The second figure shows the results of our analysis for each subject. The bars to the left represent support of the <u>active</u> enlargement hypothesis, while those to the right represent support of the <u>passive</u> enlargement hypothesis. For the first subject, the active hypothesis is supported only by the sternohyoid (77%: $p<.00003$). The active hypothesis is significantly reversed for this subject for the levator palatini (28%: $p<.00714$), which supports our hypothesis only 28% of the time. The passive hypothesis is supported for this same subject for both the constrictor (91%: $p<.00003$) and faucal pillar (71%: $p<.00135$) muscles. The active enlargement hypothesis is supported for the third subject for both the levator palatini (100%: $p<.00003$) and sternohyoid (80%: $p<.00402$), while this same subject demonstrates no significant distinction for the stop cognates for the muscles involved in passive pharyngeal enlargement. For the remaining subject, support for both hypotheses is found, the levator palatini (74%: $p<.00135$) supporting the active hypothesis and the constrictor (75%: $p<.00135$) and faucal pillar (71%: $p<.00135$) muscles supporting the passive hypothesis. The sternohyoid data for this subject are equivocal. We have also found that at least two muscles supported their hypothesis for each contrasting utterance pair for each subject.

We can see, therefore, that each subject employed a different articulatory strategy for arriving at an increased pharyngeal volume for the production

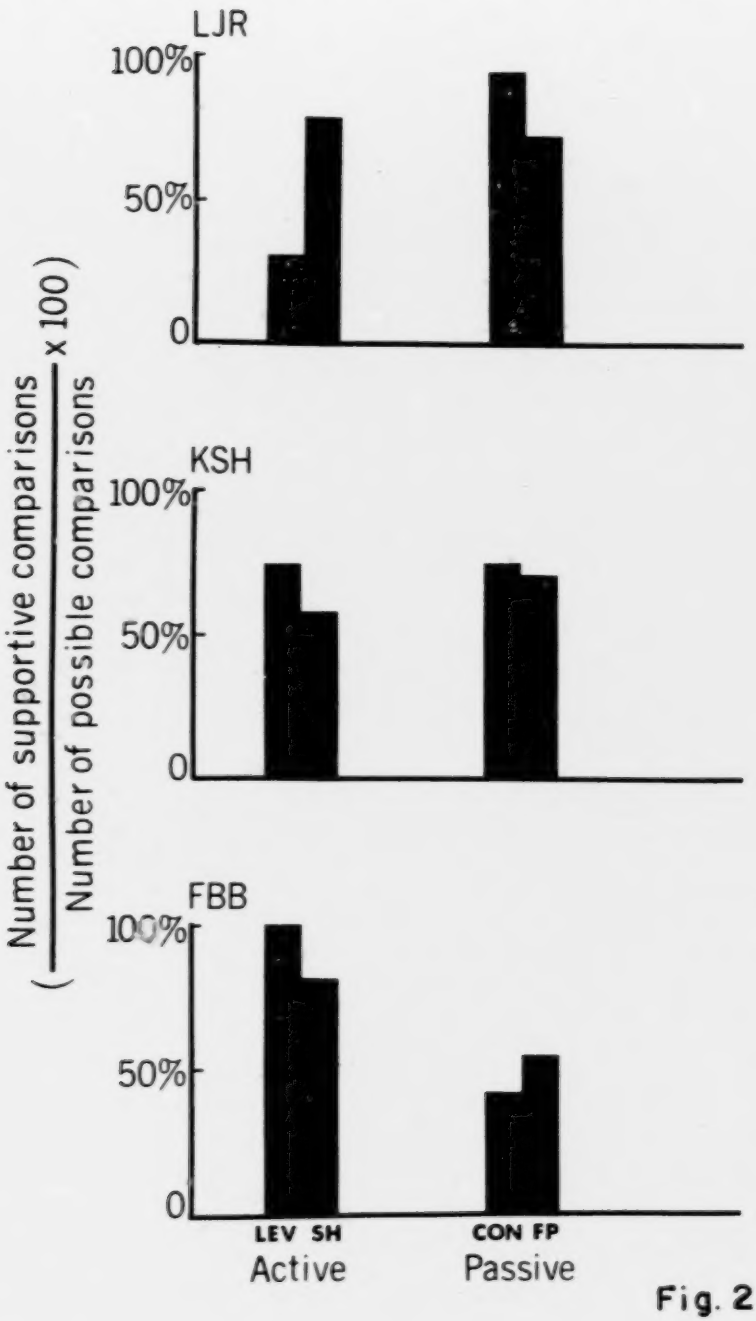

LJR
100%
50%
0
KSH
100%
50%
0
FBB
100%
50%
0
LEV SH
Active
CON FP
Passive
(Number of supportive comparisons / Number of possible comparisons x 100)

Fig. 2

of voiced stops in our sample. Our data support the suggestion of other workers that pharyngeal enlargement must in part be due to positive muscle activity: the active mode hypothesis is supported by at least one muscle for each of our subjects. The passive mode hypothesis is also supported, for two subjects, by both "passive" muscle groups. The balance of active and passive enlargement gestures, then, varies from subject to subject, although the acoustic results are equivalent in that glottal pulsing was maintained during the stop occlusion. We conclude, then, that variations in articulatory maneuvers may still result in phonetic constancy: that is, we may not specify one universal site of pharyngeal enlargement.

REFERENCES

Bell-Berti, F. and H. Hirose. (1972) Velar activity in voicing distinctions: A simultaneous fiberoptic and electromyographic study. Paper presented at the meeting of the American Speech and Hearing Association, San Francisco, Cal., November 1972. [Also in Haskins Laboratories Status Report on Speech Research SR-31/32 (this issue).]

Kent, R. D. and K. L. Moll. (1969) Vocal tract characteristics of the stop cognates. J. Acoust. Soc. Amer. 46, 1549-1555.

Perkell, J. S. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study. (Cambridge, Mass.: MIT Press).

Electromyographic Study of Speech Musculature During Lingual Nerve Block*

Gloria J. Borden,[+] Katherine S. Harris,[++] and Lorne Catena[+++]

The larger problem to which this paper is addressed is the question of the nature and use of feedback mechanisms in speech. Is skilled articulation of speech controlled by the central nervous system with little need of sensory feedback from the periphery, or is skilled speech continuously monitored by afferent information from the articulators? Experiments which we have been conducting on the speech of subjects with and without nerve blocks give weight to the theory that speech is centrally controlled once it is well learned.

We know that when a local anesthetic is applied to the inferior branch of the trigeminal nerve, there is a perceptually small effect upon the speech of some speakers. The assumption has been that blocking sensation from the tongue interferes with feedback needed for control. Last year, we presented a paper demonstrating that the nerve block traditionally used in these experiments not only affects sensory fibers, but also blocks motor nerve fibers to the mylohyoid and to the anterior digastric muscles (Borden, 1972). Since then, in an effort to clarify the results, we have attempted to produce a purely sensory block, in order to investigate its effect upon speech and upon the electromyographic signals produced by the contraction of selected muscles during speech. That is, we attempted to block the lingual nerve, which is sensory from the tongue, without affecting the motor fibers of the mylohyoid nerve.

Eleven sentences repeated eighteen times each were recorded for two subjects along with EMG recordings from the upper lip, the tongue, and certain suprahyoid muscles. These recordings were made both under normal conditions and after attempts to inject Xylocaine into the lingual nerve alone. We used complex sentences because we had previously found them to be vulnerable to the nerve block.

---

*Paper presented at the American Speech and Hearing Association Convention, San Francisco, Cal., November 1972.

[+]Haskins Laboratories, New Haven, and City College of the City University of New York.

[++]Haskins Laboratories, New Haven, and Graduate Division of the City University of New York.

[+++]Southern Illinois University, Edwardsville.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

The results indicate, first, that the nerve block had a rather dramatic effect on the contraction of the intrinsic tongue muscles from which we recorded. Figure 1 shows the activity of the superior longitudinal muscle. The darker line is always the activity during the nerve block condition. This subject evidenced a significant drop in activity in the right superior longitudinal (SL) muscle while the anesthesia was in effect. Normally, the SL peaks for /θ/, as in this utterance "It could be the thirsty wasp." The electrode placed on the left side showed a similar decrease in activity.

A second subject, however, reacted quite differently to the nerve block. Figure 2 shows the utterance "It could be the school blocks" for the second subject. The SL which normally peaks for /l/ can be seen to have been affected by the block on the right side in a way similar to the first subject. The left electrode, in contrast, recorded much more electrical activity during the block condition than during the normal condition. We do, then, see that the block has an effect upon these tongue muscles. In the sense that the lingual nerve is sensory from the anterior part of the tongue as a whole, we classify these muscles as associated with sensory fibers from the blocked branch of the trigeminal nerve. The effect of the nerve block was general depression of activity in one subject and in the other subject, one side depressed, while the other side evidenced greater effort under nerve block.

We turn now to the effect of this block upon muscles served not by sensory nerves involved in this nerve block but by motor nerves. Figure 3 shows the recording from the mylohyoid muscle for the utterance "It could be the cat's whiskers." This muscle which normally contracts for the stop consonant /k/ seems to be totally blocked on the right side and partially blocked on the left side. Thus, it seems that despite our efforts to avoid infiltration of the anesthesia down to the mylohyoid nerve, we were getting a significant effect if not a complete paralysis of the mylohyoid muscle, indicating an effective nerve block of that branch. The other subject showed the same motor effect. The inactivity of the mylohyoid muscle was consistent at both left and right placements and held for each token of each utterance type.

The other muscle served by motor fibers from the anesthetized nerve is the anterior digastric, its motor fibers being the lowest branch of the trigeminal nerve. Again, both subjects show a depressed anterior digastric during the nerve block condition.

So far, we have seen decided changes during nerve block, some of which could be explained on the basis of sensory interference with motor control and some of which could be explained on the basis of a loss of motor nerve function. Let us look now at the muscles which we have called "other," because they are muscles which are not served by fibers of the mandibular branch of the trigeminal nerve. There is, therefore, no direct basis for change; that is, the nerve fibers innervating these muscles should not have been directly affected by the anesthesia. Figure 4 shows the activity of the superior orbicularis oris with the electrode placed just left of the midline on the upper lip. It normally peaks for labial closure, as for /b/'s and /p/'s in "It could be the spring grapes." For the first subject it is somewhat depressed in amplitude during nerve block, but for the second subject, it is much more active than normal.

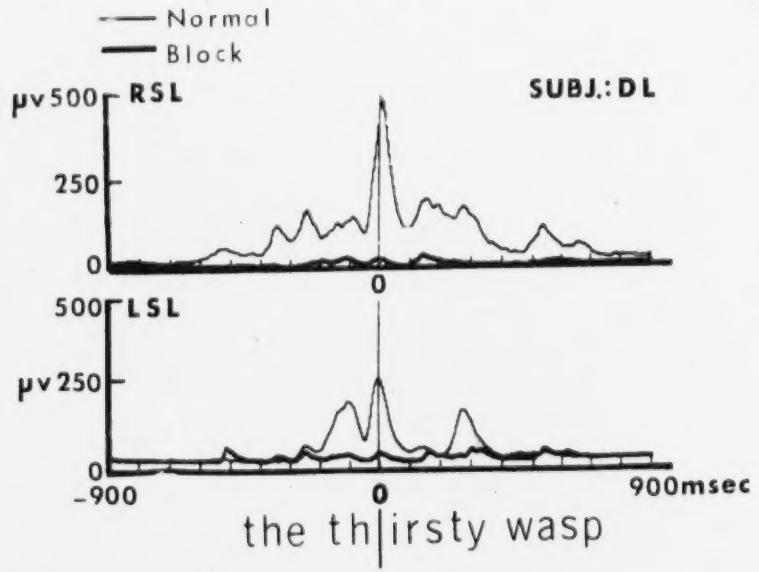

Normal
Block
μv500
RSL
SUBJ.: DL
250
0
0
500
LSL
μv250
0
-900
0
900msec
the th|irsty wasp

Fig. 1

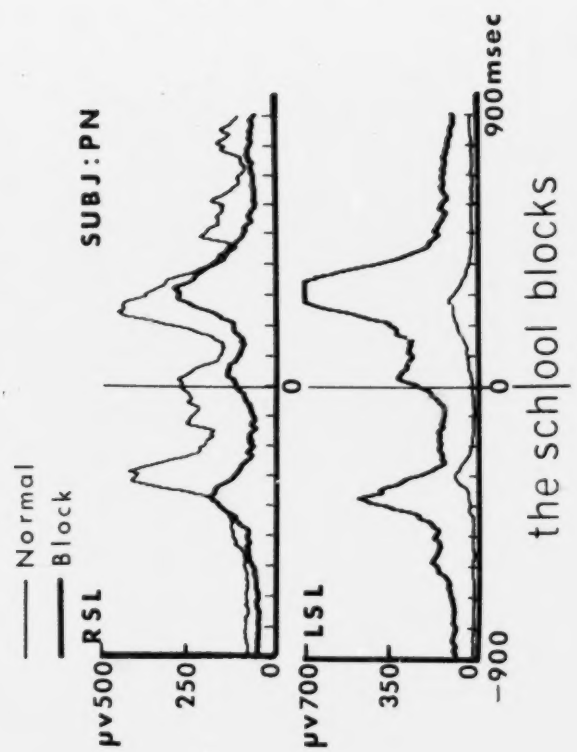

Normal
Block
SUBJ:PN
RSL
μv500
250
0
LSL
μv700
350
0
0
0
−900
900msec
the school blocks

Fig. 2

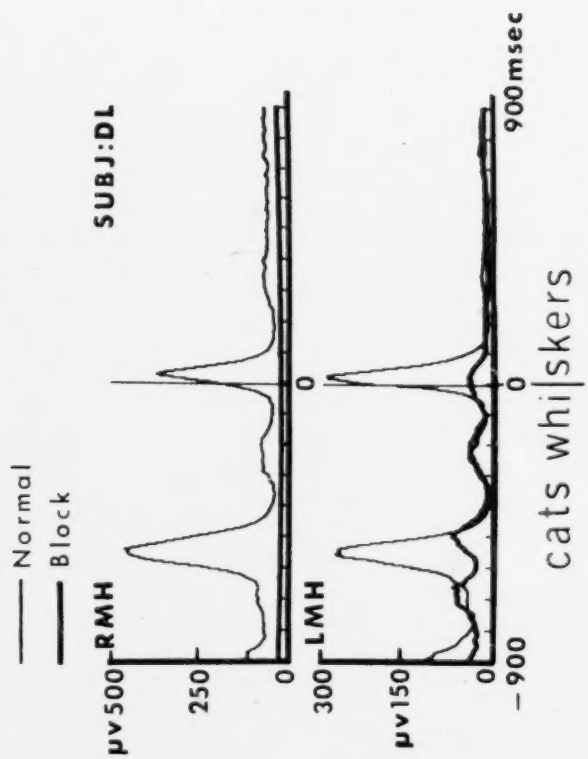

Normal
Block
SUBJ:DL
RMH
μv500
250
0
LMH
300
μv150
0
0
−900
900msec
cats whi|skers

Fig. 3

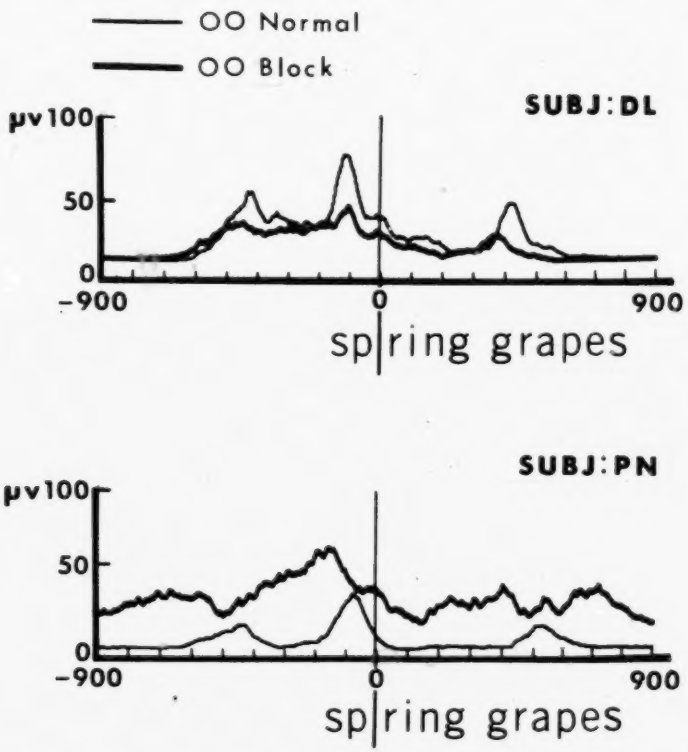

OO Normal
OO Block
SUBJ:DL
μv100
50
0
-900
0
900
sp|ring grapes
SUBJ:PN
μv100
50
0
-900
0
900
sp|ring grapes

Fig. 4

Another muscle which is not known to have either motor or sensory innervation from the blocked nerve is the genioglossus. It showed the same changes in amplitude under the blocked condition as did the orbicularis oris. These effects upon what we have called the other muscles cannot be explained by any direct peripheral effect, since they are apparently not served by either motor or sensory nerves affected by this nerve block.

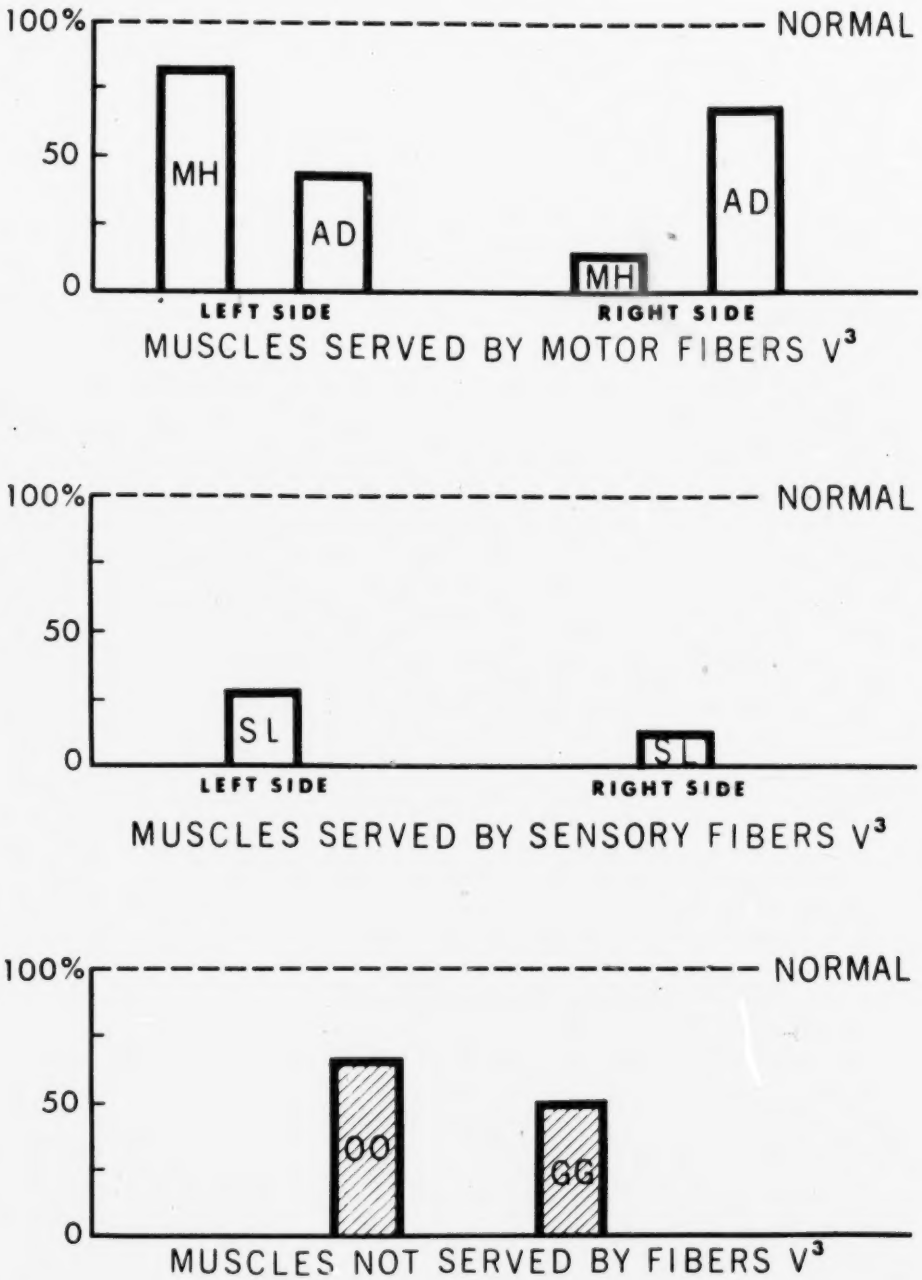
The examples just shown are typical of the effects for each muscle for all eleven utterance types. When the absolute peak values in microvolts during nerve block are compared to the normal peak values, and the percent of normal is averaged for each muscle, we can see the pooled difference from the normal condition which the nerve block produces in one subject, shown in Figure 5. The nerve block produces a consistently depressed state of activity in this subject. The general depression in muscles extends even to those muscles which should be completely unaffected by the block and therefore not explained on the basis of gamma efferent system interference or simple motor or sensory information loss.

The other subject, Figure 6, although apparently suffering motoric denervation and some direct sensory effect, seems to be performing compensatory activity in an attempt to counteract the direct physical effects of the block. To the extent that we could generalize these results, there would seem to be individual reactions to oral anesthesia, general depression in some speakers and increased effort in other speakers, presumably part of a compensatory struggle.

To summarize the effects of the nerve block in this experiment, the first class of muscles, those innervated by motor fibers from the blocked nerve were consistently depressed or inactive. The next two classes of muscles, those presumably served by sensory fibers from the blocked nerve, and those which should be independent of the blocked nerve were sometimes less active, sometimes more active, depending upon the side of electrode placement and upon the idiosyncratic reaction of the subject.

If we may hypothesize a bit on the basis of the 11 subjects we have studied so far (7 with nerve block without EMG recordings, and 4 with nerve block on whom we have EMG recordings), it is apparent that we are getting effects which cannot be explained peripherally. We hypothesize two theories to account for the muscle behavior which we have noticed. One idea is that these effects are due to compensation reorganization, that is, some people, with the realization that there is a loss of sensation and that their speech is slurred, will attempt to reduce the speech distortion by increasing activity in some speech muscles. The other possibility is that we are seeing, in addition to any peripheral effects such as the inactive mylohyoid muscles and the depressed intrinsic tongue muscles, an additional effect which is not peripheral but a generalized depression of central activity. Drowsiness after Xylocaine is a well known side effect. Pharmaceutical studies indicate that local anesthesia may appear in considerable quantities in the blood stream (de Jong, 1968), and a speech effect is one clinical sign of a rising level of anesthesia in the blood. Furthermore, it has been shown that local anesthetics readily cross the blood-brain barrier (Usubiaga, Moya, Wikinski, and Usubiaga, 1967). It is possible that a slight loss of central control may relate more directly to the slurring of speech than either the motor or sensory effects which we are witnessing at the periphery. Of course, this idea is entirely speculative, but it is a possibility which we intend to pursue. The speech effect, when it does exist

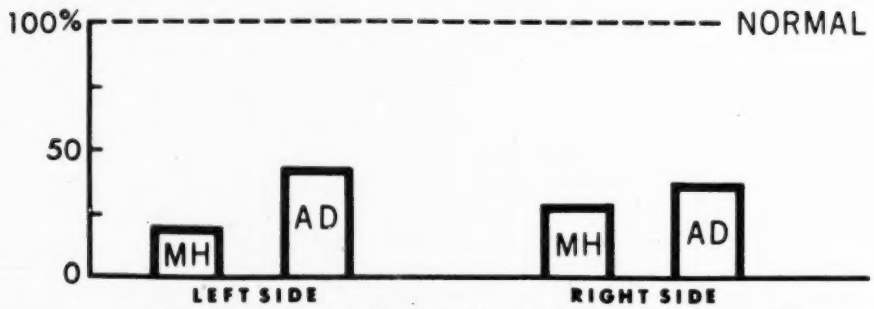
SUBJECT : DL



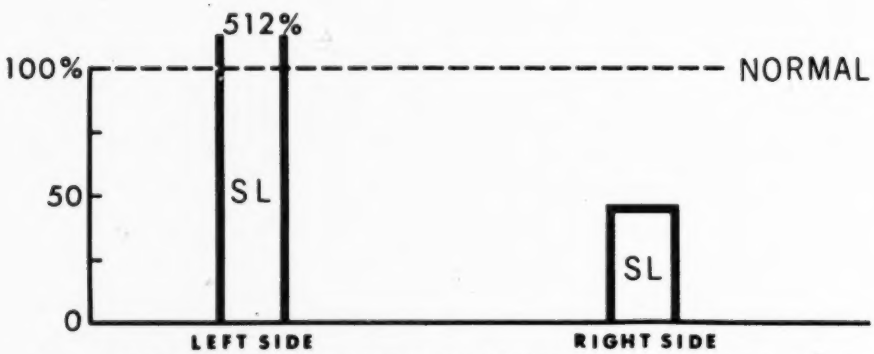
Mean percentages of normal peak EMG amplitudes in microvolts for muscles during nerve block.

Fig. 5

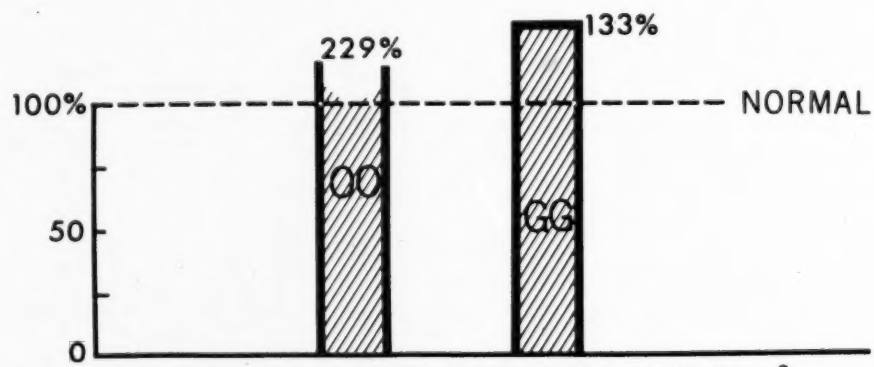
SUBJECT: PN



MUSCLES SERVED BY MOTOR FIBERS V^3



MUSCLES SERVED BY SENSORY FIBERS V^3



MUSCLES NOT SERVED BY FIBERS V^3

Mean percentages of normal peak EMG amplitudes in microvolts for muscles during nerve block.

Fig. 6

[and it often does not exist (Borden, 1971)], sounds perceptually very like 'drunk' speech. We accept 'drunk' speech to be a consequence of the alcohol having crossed the blood-brain barrier to affect the central control of speech.

One final point which gives weight to the 'drunk speech' hypothesis is that the speech sounds most noticeably affected by the mandibular nerve block are sounds which apparently rely more upon proprioceptive than on tactile information. A person learning the /s/ and /r/ gestures probably relies upon some sense of tongue movement and position to correlate with the acoustic information. It had long been assumed that the lingual nerve carried proprioceptive as well as tactile information from the anterior 2/3 of the tongue, because the hypoglossal nerve has no sensory root. Studies by Bowman and Combs (1968) would indicate that nerve fibers from muscle spindles in the tongue do course along the hypoglossal nerve for part of the way and then cross to join some cervical nerves in rhesus monkeys. If this is the case in humans, then the lingual nerve would participate in relaying only tactile sensation and we would expect anesthesia of this nerve to affect such sounds as /t/, /d/, and /n/. The fact that these consonants remain relatively undistorted forces us to look elsewhere for an explanation of the speech effect. Looking higher up in the nervous system seems reasonable.

In conclusion, the still unanswered question of whether skilled speech is centrally controlled without need of feedback from the periphery or whether it is monitored as we speak by afferent information from the articulators remains an interesting problem and one worthy of further investigation.

REFERENCES

Borden, G. J. (1971) Some effects of oral anesthesia upon speech: A perceptual and electromyographic analysis. Ph.D. dissertation, City University of New York.

Borden, G. J. (1972) Some effects of oral anesthesia upon speech: An electromyographic investigation. Haskins Laboratories Status Report on Speech Research SR-29/30, 27-47.

Bowman, J. P. and C. M. Combs. (1968) The discharge patterns of lingual spindle afferent fibers in the hypoglossal nerve of the rhesus monkey. Exp. Neurol. 21, 105.

de Jong, R. H. (1968) Local Anesthetic seizures. Surg. Dig. 3, 30.

Usubiaga, J. E., F. Moya, J. A. Wikinski, and L. E. Usubiaga. (1967) Relationship between the passage of local anesthetics across the blood-brain barrier and their effects on the central nervous system. Brit. J. of Anesth. 39, 943-946.

Velar Activity in Voicing Distinctions: A Simultaneous Fiberoptic and Electromyographic Study*

Fredericka Bell-Berti[+] and Hajime Hirose[++]
Haskins Laboratories, New Haven

PURPOSE

The present study was undertaken in order to determine the relationship between increases in electromyographic potential and articulator movements, the articulator in this case being the soft palate. Our immediate aim was to provide simultaneous measures of velar height and EMG potential which would strengthen our belief that different levels of EMG activity in minimal contrasts may be used as an indicator of differences in articulator position. A year ago (Berti and Hirose, 1971) we reported differences in the magnitude of EMG signals recorded from the muscles of the velopharyngeal region for the production of voiced and voiceless stop consonants. We interpreted these differences as indicating differences in the magnitude of articulator displacement. Greater EMG activity in the levator palatini for voiced stops was taken to mean an increased velar height, hence, increased pharyngeal volumes for voiced stops than for their voiceless cognates. We assumed that the levator palatini is the muscle essentially responsible for palatal elevation, and that we were then dealing with a simple one muscle-one parameter system (although we know that contraction of the levator palatini results in movements of the soft palate in more than just the vertical dimensions--that is, it also moves the palate posteriorly).

METHODS

A subset of our original stimulus inventory was used. We compared voiced and voiceless labial stop consonants in nasal-oral (fimbip) and oral-nasal (fibmip) contrasts within the vowel environments /i/ and /ɑ/. Hooked-wire electrodes were inserted perorally into the dimple of the soft palate. The EMG potentials were recorded into magnetic tape. To assist

---

*Paper presented at the American Speech and Hearing Association Convention, San Francisco, Cal., November 1972.

[+]Also the Graduate School of the City University of New York and Montclair State College, Upper Montclair, N. J.

[++]Also Faculty of Medicine, University of Tokyo.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

with identification of palatal movement, a grid made of a thin plastic film was placed along the floor of the nasal cavity. A fiberoptic endoscope was inserted to the subject's right nostril and positioned to provide a view of the velum as it was raised and lowered. A random list of our eight utterance types was repeated ten times. Motion pictures were taken through the fiberscope, at 60 frames/sec, of all repetitions of the utterance list. A synchronization mark was recorded on the EMG data tape. The line-up point chosen for averaging tokens was the end of /m/ when it preceded the medial stop (for example, at the end of the /m/ in /fimbip/), or the beginning of /m/ when it followed the stop (for example, the onset of /m/ in /fibmip/). This point is identified as "zero" on the abscissa. The EMG potentials were rectified, integrated, and computer-averaged for eight to ten tokens of each utterance type. Frame-by-frame measurements of velar height were also made for each of eight to ten tokens of each utterance type. Only vertical movement of the soft palate was determined. The velar height measurements were averaged for each token of each utterance type.

RESULTS

Figure 1

Comparisons of EMG activity and velar height reveal timing differences between the two measures, as we would expect. Increases in EMG activity always precede velar elevation. The upper figure shows that peak EMG activity for an utterance in which /m/ precedes /b/ (/fimbip/) occurs 10 msec prior to the end of /m/, while peak velar height occurs approximately 100 msec after the end of /m/, a temporal separation of approximately 110 msec. The lower figure which compares EMG activity and velar height for an utterance in which /b/ follows an oral vowel (/fibmip/) demonstrates a temporal separation of the peaks of about 60 msec, with the EMG peak again preceding peak velar height. The time lag between the EMG peak and peak velar height is generally greater for a stop following a nasal (/m/), in this case about 110 msec, than for the stop following an oral vowel (/i/), in this case about 60 msec.

Figure 2

In addition to differences in the time-lag between peak EMG activity and peak velar height for stop consonants in these environments there are also differences in the magnitude of velar movements and their corresponding EMG potentials. Inspection of the EMG potentials for these utterances (the lower figure) reveals a greater increase in activity, as well as a greater peak magnitude, of the EMG potential for the /b/ which follows the /m/ (in /fimbip/) than for the /b/ which follows the /i/ (in /fibmip/). The solid line in the upper figure, representing velar height for the utterance in which /b/ follows /i/ (/fibmip/) remains above the dashed line until after the "zero" line-up point. The peak velar height associated with the /b/ articulation (at 4 frames before zero) for this utterance, in which /b/ follows /i/ (/fibmip/), is greater than the corresponding peak (at 6 frames after zero) for the utterance in which /b/ follows /m/ (/fimbip/). Although the increase in velar height is greater when /b/ follows /m/ (/fimbip/), absolute velar height may not reflect this difference.

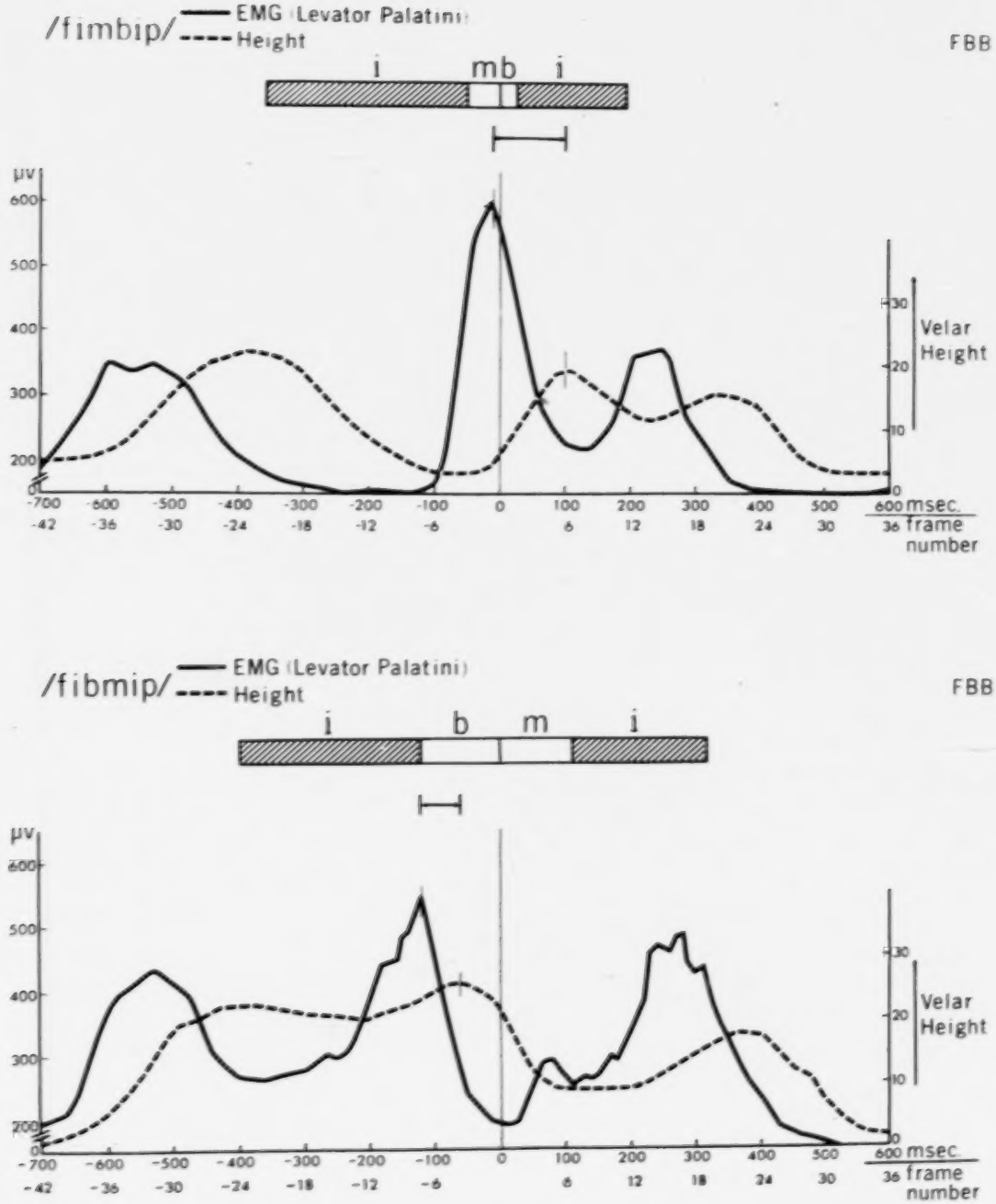
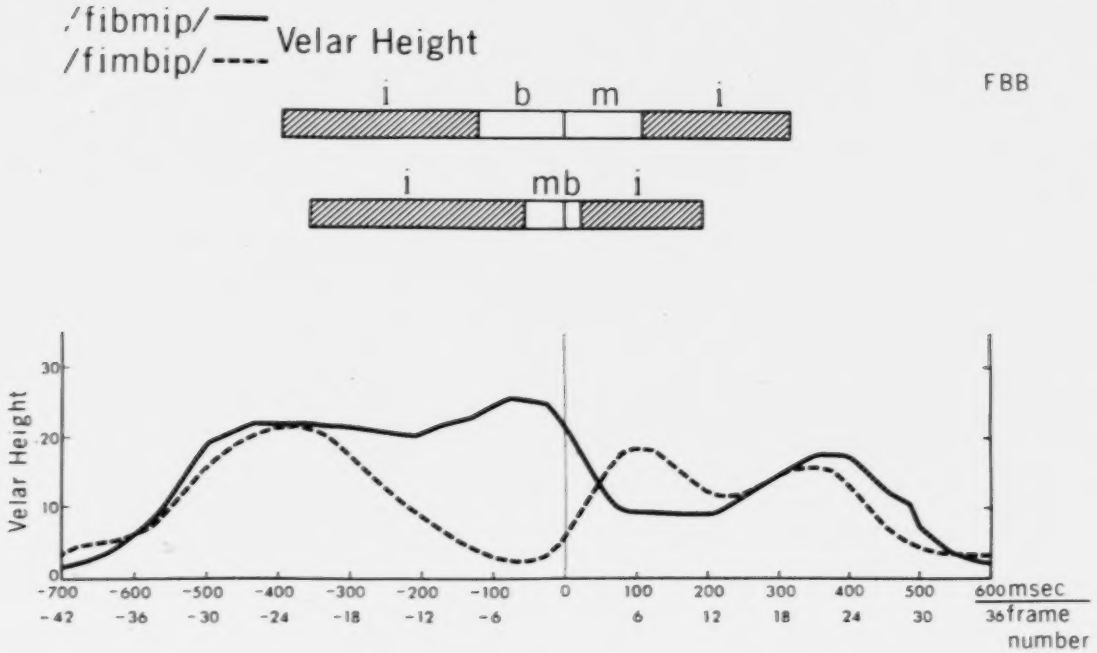

/fimbip/
EMG (Levator Palatini)
Height
FBB
i mb i
μV
600
500
400
300
200
0
-700 -600 -500 -400 -300 -200 -100 0 100 200 300 400 500 600 msec.
-42 -36 -30 -24 -18 -12 -6 6 12 18 24 30 36 frame number
30
20
10
0
Velar Height
/fibmip/
EMG (Levator Palatini)
Height
FBB
i b m i
μV
600
500
400
300
200
0
-700 -600 -500 -400 -300 -200 -100 0 100 200 300 400 500 600 msec.
-42 -36 -30 -24 -18 -12 -6 6 12 18 24 30 36 frame number
30
20
10
0
Velar Height

Fig. 1


/fibmip/
/fimbip/
Velar Height
FBB
i b m i
i mb i
Velar Height
30
20
10
0
-700 -600 -500 -400 -300 -200 -100 0 100 200 300 400 500 600 msec
-42 -36 -30 -24 -18 -12 -6 6 12 18 24 30 36 frame number

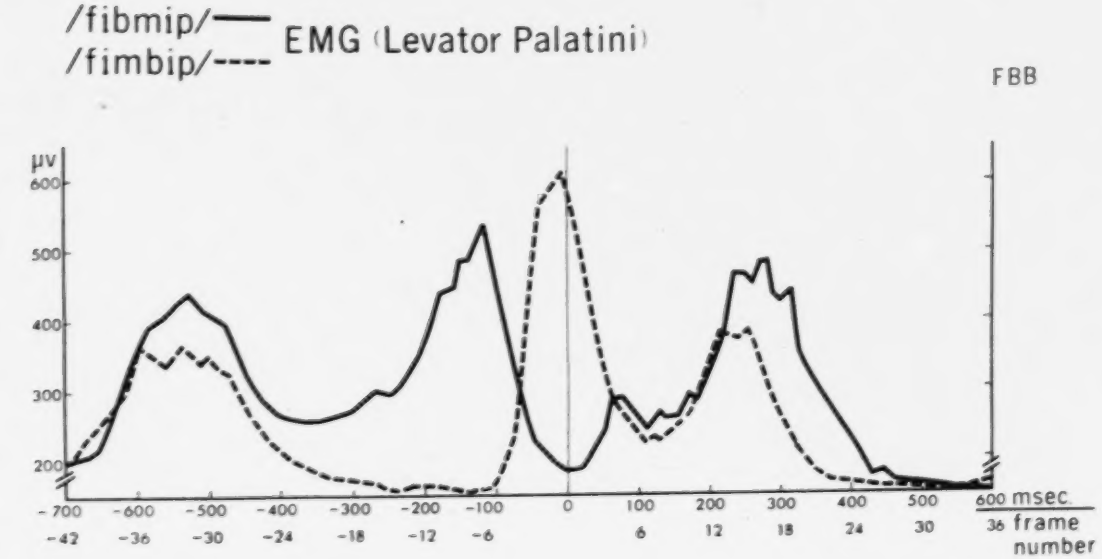

/fibmip/
/fimbip/
EMG (Levator Palatini)
FBB
μv
600
500
400
300
200
-700 -600 -500 -400 -300 -200 -100 0 100 200 300 400 500 600 msec
-42 -36 -30 -24 -18 -12 -6 6 12 18 24 30 36 frame number

Fig. 2

	1 EMG INCREASE	2 HEIGHT INCREASE	3 BASELINE HEIGHT	4 TIME LAG
ORAL-NASAL	177. μv	4.	13.5	95 MSEC
NASAL-ORAL	348. μv	13.5	1.5	130 MSEC

Fig. 3

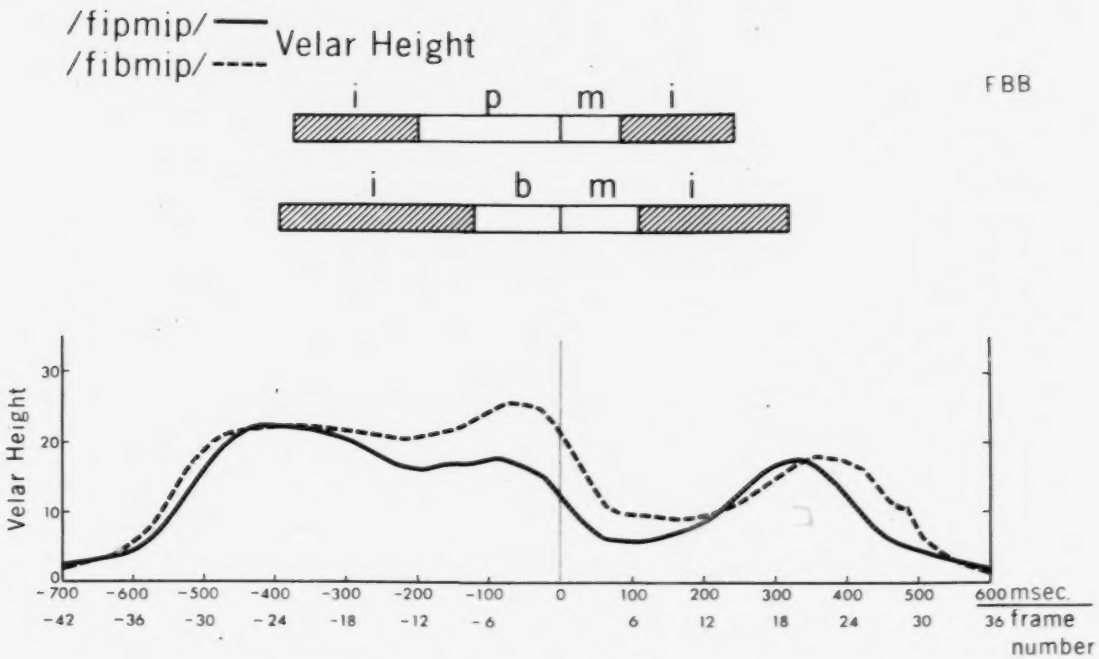
Figure 3

When increases in velar height and the EMG potentials for stop consonants in all oral-nasal contrasts (where the stop follows a vowel) are pooled and compared with all nasal-oral contrasts, three trends are revealed. Looking at Column 1, we see that there is a greater average increase in EMG activity for stop consonant articulation in nasal-oral utterances (348.μv) than in oral-nasal utterances (177.μv). The second trend is revealed in Column 2: there is a greater increase in velar height for stop consonants in nasal-oral utterances (13.5 units) than for stops in oral-nasal utterances, that is, where the stop follows a vowel (4. units), even though absolute velar height may be greater for a stop in an oral-nasal contrast as shown in Figure 2. This is possible because the starting point against which the increase in velar height was measured, which is given in Column 3, is considerably higher for oral-nasal than for nasal-oral utterances. The third distinction between stops in the two environments is that the average time-lag between peak EMG activity for a stop consonant and peak velar height for that stop is greater when the stop follows a nasal (approximately 130 msec) than when the stop follows a vowel (approximately 95 msec). This might be explained as a function of the increased work-load when the palate must be elevated from a lower position.

A Pearson product moment correlation coefficient was computed between our two parameters: that is, for the increases in velar height and the increases in EMG potential. The Pearson r_{xy} is .84, which is significant at the 0.005 level. This implies that given equal starting points, the greater the EMG activity the greater the velar height.

CONCLUSION

In conclusion, although there is no obvious constant relationship between absolute velar height and the absolute magnitude of the EMG signal for that articulation, there is a strong correlation between the magnitude of the increase in EMG potential and the magnitude of the change in velar height. That is, the larger the increase in EMG potential, the greater will be the corresponding increase in velar height. This conclusion is supported by the data displayed in Figure 4 for the minimal utterance pair /fipmip-fibmip/. The EMG potentials for the stop consonant articulation are in the


/fipmip/
/fibmip/
Velar Height
FBB
i p m i
i b m i
Velar Height
30
20
10
0
-700 -600 -500 -400 -300 -200 -100 0 100 200 300 400 500 600 msec.
-42 -36 -30 -24 -18 -12 -6 6 12 18 24 30 36 frame number

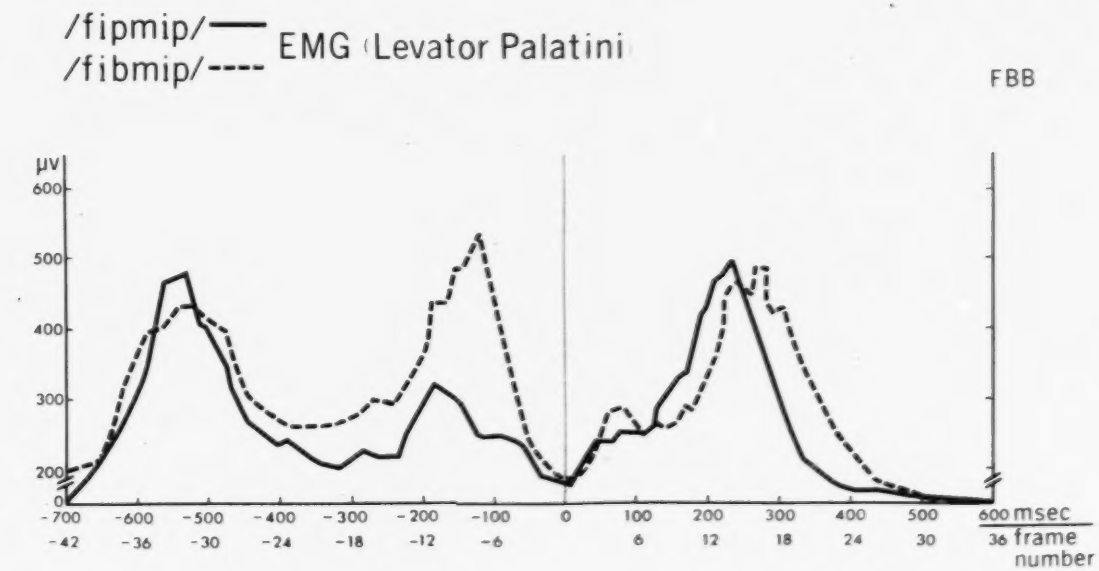

/fipmip/
/fibmip/
EMG (Levator Palatini)
FBB
μv
600
500
400
300
200
0
-700 -600 -500 -400 -300 -200 -100 0 100 200 300 400 500 600 msec
-42 -36 -30 -24 -18 -12 -6 6 12 18 24 30 36 frame number

Fig. 4

lower figure. The EMG peak is greater for /b/ (the dashed line) than /p/ (the solid line) in utterances in which both follow the same vowel. The upper figure reflects this difference in the level of EMG activity: the curve of velar height is higher for the /b/ than for the /p/ in the region of the voiced-voiceless comparison, while the velar height curves are similar for those regions in which the EMG curves are similar.

We feel, therefore, that for minimal utterance pairs increases in EMG potential should be interpreted as reflecting increases in velar height. Our earlier results, in which subjects varied in the differences they demonstrated in EMG potentials for voiced and voiceless stop consonants (in similar phonetic environments) may be interpreted as reflecting differences in velar height, and, therefore in pharyngeal cavity size.

SUMMARY

In summary, we compared simultaneous measures of velar height and EMG activity of the levator palatini for utterances contrasting voiced and voiceless consonants in various environments. We found a strong correlation between the size of the increase in the EMG signal and the size of the increase in articulator displacement. We concluded, therefore, that differences in the strength of EMG signals for contrasting phonemes in otherwise constant environments should be interpreted as differences in velar height.

REFERENCE

Berti, F. B. and H. Hirose. (1971) Velopharyngeal function in oral/nasal articulation and voicing gestures. Haskins Laboratories Status Report on Speech Research SR-28, 143-156.

Electromyographic Study of the Tones of Thai*

Donna Erickson[+] and Arthur S. Abramson[+]
Haskins Laboratories, New Haven

To study the correlation of laryngeal muscle activity with the rather complex use of fundamental frequency changes in a tone language, we have done an electromyographic analysis of the production of the tones of Standard Thai. The Thai language has five tones, which are called mid, high, low, rising, and falling. The primary physical correlate of each tone--though not necessarily the only one (Abramson, in press)--is a movement of fundamental frequency through a typical but relative contour on the syllable regardless of the speaker's voice range or indeed the choice of the speaker (Abramson, 1962; Chiang, 1967; Howie, 1970). The f_0 contours are determined by variations in the rate of laryngeal vibration, which is regulated to a considerable degree by the laryngeal muscles (Sawashima, in press) as well as by subglottal air pressure.

Our primary purpose is to examine the laryngeal mechanisms underlying the phonemic tones of Thai to see whether the abstract contours isolated for them by Abramson (1962) are well specified as typical patterns of muscle activity. In addition, some of the questions that have arisen over the use of certain laryngeal muscles (Gårding, 1970; Simada and Hirose, 1971; Ohala, 1972) can be profitably examined by looking at a tone language in which seemingly heavy demands are placed upon the speaker to produce a lexically appropriate tonal contour for every syllable.

For this study we had four native speakers of Thai undergo hooked wire electrode placement in five laryngeal muscles: cricothyroid, vocalis, sternohyoid, sternothyroid, and thyrohyoid. The insertions were made by Hajime Hirose (Hirose, 1971), and the electromyographic data were processed by computer averaging techniques (Port, 1971). The fundamental frequency contours were derived from the Pitch Period Extractor, developed by G. Fant and J. Liljencrants and made available to us by George Allen of the Dental Research Center of the University of North Carolina. Each speaker said 45 short sentences. Included in this list were the five tones in nine contexts: the three long vowels /aa ii uu/ and the three labial consonants /b p ph/. These three consonants were chosen as representative of the three degrees of voicing in the language. It was felt that these vocalic and consonantal environments might affect the f_0 contours and bring about differences in the patterns of laryngeal muscle activity.

---

*Paper presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., November 1972.

[+]Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-31/32 (1972)]

RISING TONE

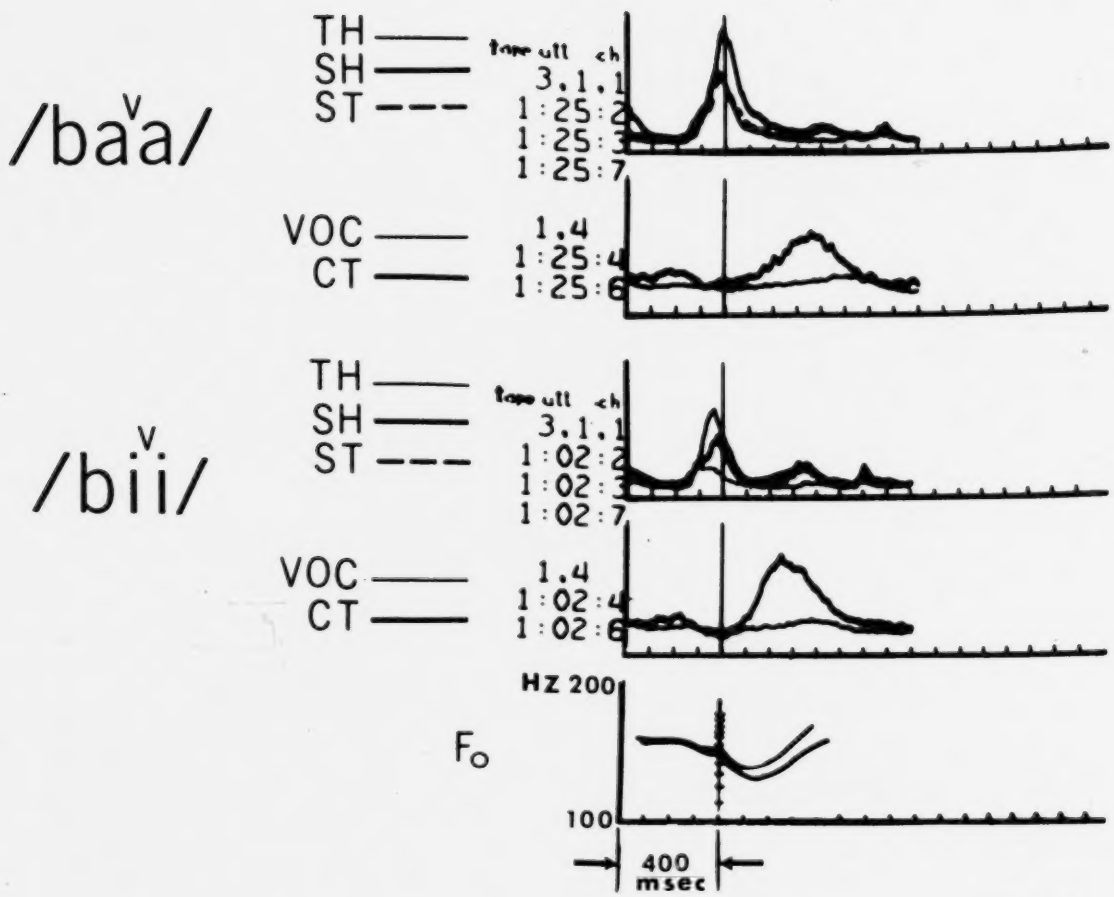


Fig. 1

FALLING TONE

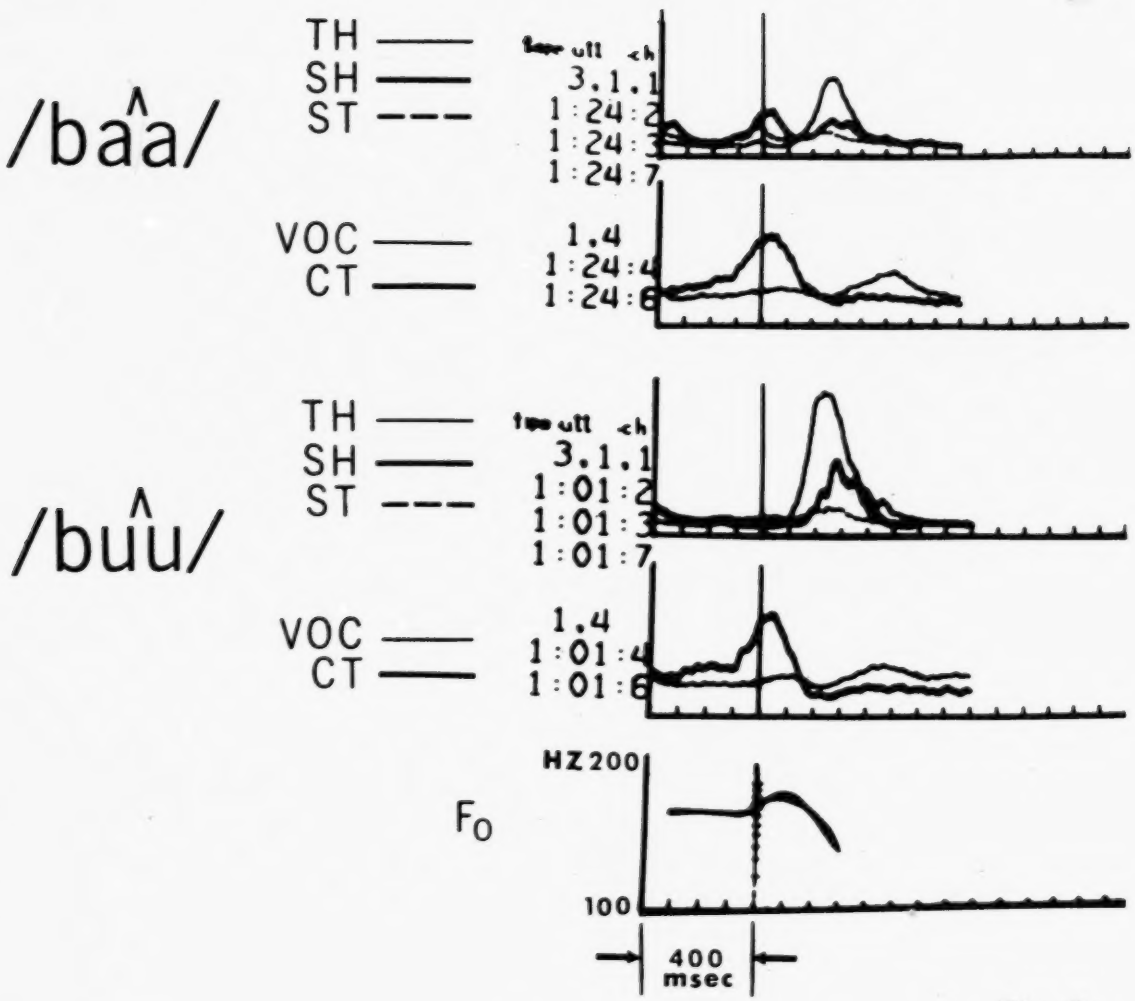


Fig. 2

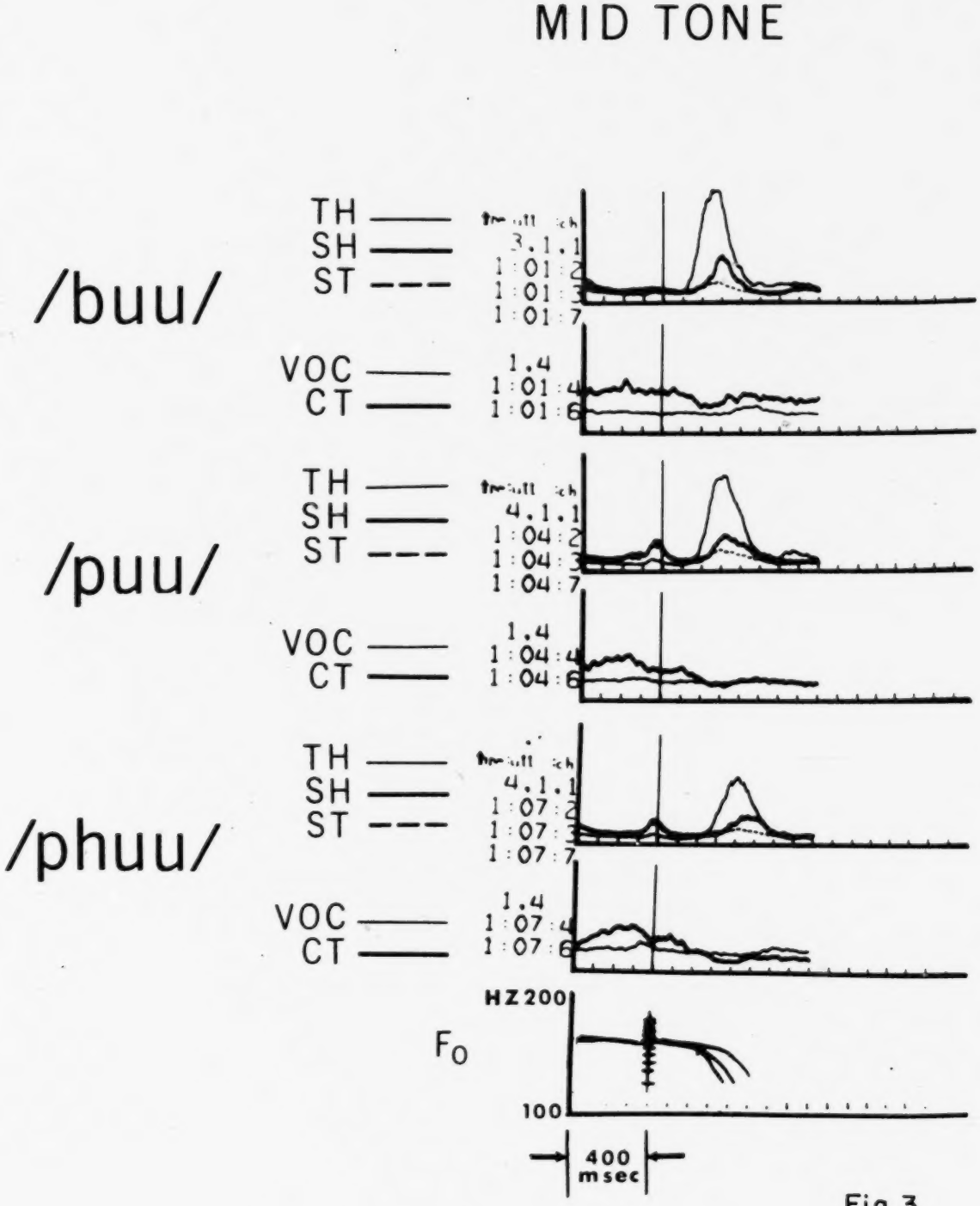

MID TONE
/buu/
/puu/
/phuu/
TH
SH
ST
VOC
CT
3.1.1
1:01:2
1:01:3
1:01:7
1.4
1:01:4
1:01:6
4.1.1
1:04:2
1:04:3
1:04:7
1:04:4
1:04:6
1:07:2
1:07:3
1:07:7
1:07:4
1:07:6
HZ200
Fo
100
400
msec

Fig. 3

We generally find that the activity of the cricothyroid muscle increases with the raising of f_0, and the activity of the strap muscles--the sternohyoid, the sternothyroid, and the thyrohyoid--increases with the lowering of f_0. For an illustration of this see Figure 1, which displays the muscle activity and concomitant f_0 contour of the syllables /bǎa/ and /bǐi/ on the rising tone. In this figure and the following two, each tracing is an average of approximately 16 utterances produced by one of the Thai subjects. /bǎa/ has the lower of the two f_0 contours. The vertical line indicates the release of the stop. Notice that for the rising tone there is an initial increase in activity of the extrinsic muscles at the release of the stop, which corresponds to an initial drop in f_0 of the rising tone; the increase in activity of the cricothyroid muscle about 200 msec after the release corresponds to a subsequent rise in f_0. The vocalis in the case of this speaker does not seem to be particularly active with respect to changes in f_0. Figure 2 shows the same kind of activity in connection with f_0 contours of the falling tone.

For the effects of differing vocalic environments, see Figure 2 for a display of the utterances /bâa/ and /bûu/ on the falling tone. The f_0 contour of /bâa/ is the lower of the two f_0 contours. Note that there is activity of the extrinsic muscles at the release of the stop for /bâa/, but not for /bûu/. This increased activity of the extrinsic muscles for the vowel /aa/ must be the effect of jaw opening for this vowel. The effect can also be observed in Figure 1.

The effects of the voicing distinctions in the initial consonants are less clear. Some contrast in extrinsic muscle activity between the pre-voiced /b/ consonant and the /p/ and /ph/ consonants can be seen in Figure 3, a display of the mid tone as produced in the utterances /buu puu phuu/. For the two voiceless consonants /p ph/ there is a small peak of activity just before the release of the stop, but none for the voiced stop /b/. The longest of the f_0 contours is for /ph/, the shortest is for /b/. We have not yet analyzed the data for the expected differences in f_0 linked to the voicing states of the initial consonants, and, beyond that, possible correlations between these and details of the muscle activity.

For the most part all four speakers use their muscles in much the same way. One cross-speaker difference seems to be the extent to which the vocalis muscle is active during f_0 rises. This may be a secondary effect of anatomical differences among the speakers.

Although portions of the f_0 contours must be controlled by variations in subglottal air pressure, we have found a characteristic pattern of laryngeal muscle activity for each of the distinctive tones of Thai. We have yet to explore more thoroughly the effects of contextual factors.

REFERENCES

Abramson, A. S. (1962) The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments. (Bloomington: Indiana University Research Center in Anthroprology, Folklore, and Linguistics, Pub. 20).

Abramson, A. S. (in press) Tonal experiments with whispered Thai. In Papers on Linguistics and Phonetics in Memory of Pierre Delattre, ed. by A. Valdman. (The Hague: Mouton). (Also in Haskins Laboratories Status Report on Speech Research SR-19/20, 37-57.)

Chiang, H. T. (1967) Amoy-Chinese tones. Phonetica 17, 100-115.
Gårding, E. (1970) Word tones and larynx muscles. Working Papers, Phonetics Laboratory, Lund University 3, 20-45.
Hirose, H. (1971) Electromyography of the articulatory muscles; current instrumentation and technique. Haskins Laboratories Status Report on Speech Research SR-25/26, 73-86.
Howie, J. M. (1970) The vowels and tones of Mandarin Chinese: Acoustical measurements and experiments. Indiana University Ph.D. dissertation.
Ohala, J. J. (1972) How is pitch lowered? J. Acoust. Soc. Amer. 52, 124(A).
Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.
Sawashima, M. (in press) Laryngeal research in experimental phonetics. In Current Trends in Linguistics 12, ed. by Thomas A. Sebeok et al. (The Hague: Mouton). (Also in Haskins Laboratories Status Report on Speech Research SR-23, 69-115.)
Simada, Z. and H. Hirose. (1971) Physiological correlates of Japanese accent pattern. Research Institute of Logopedics and Phoniatrics, University of Tokyo 5, 41-50.

II. PUBLICATIONS AND REPORTS

III. APPENDICES

PUBLICATIONS AND REPORTS

Publications and Manuscripts

Auditory and Phonetic Processes in Speech Perception: Evidence from a Dichotic Study. M. Studdert-Kennedy, D. Shankweiler, and D. Pisoni. Cognitive Psychology (1972) 3, 455-466.

The Activity of the Intrinsic Laryngeal Muscles in Voicing Control: An Electromyographic Study. H. Hirose and T. Gay. Phonetica (1972) 25, 3, 140-164.

On Stops and Gemination in Tamil. Leigh Lisker. International Journal of Dravidian Linguistics (1972) 1, 144-150.

*Machines and Speech. Franklin S. Cooper. In Research Trends in Computational Linguistics, report of a conference, 14-16 March 1972.

Language Codes and Memory Codes. A. L. Liberman, I. G. Mattingly, and M. T. Turvey. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington: V. H. Winston, 1972) 307-334.

The following four papers appeared in Language by Ear and by Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and Ignatius G. Mattingly. (Cambridge, Mass.: MIT Press, 1972):

How is Language Conveyed by Speech?
Franklin S. Cooper, 25-45.

Reading, the Linguistic Process, and Linguistic Awareness.
Ignatius G. Mattingly, 133-147.

Misreading: A Search for Causes.
Donald Shankweiler and Isabelle Y. Liberman, 293-318.

Background of the Conference.
J. J. Jenkins and Alvin M. Liberman, 1-2.

Laryngeal Control in Vocal Attack: An Electromyographic Study. H. Hirose and T. Gay. Folia Phoniatrica, in press. (Also in SR-29/30, 1972.)

Voicing-Timing Perception in Spanish Word-Initial Stops. A. S. Abramson and Leigh Lisker. Journal of Phonetica (1973) 1, 1, in press.

*Constructive Theory, Perceptual Systems, and Tacit Knowledge. M. T. Turvey. In Cognition and Symbolic Activity, ed. by D. Palermo and W. Weimer. (Washington, D. C.: V. H. Winston, in press). Proceedings of the Conference on Cognition and the Symbolic Processes, Pennsylvania State University, October 1972.

---

*Appears in this report, SR-31/32.

*Effect of Speaking Rate on Labial Consonant Production: A Combined Electromyographic-High Speed Motion Picture Study. Thomas Gay and Hajime Hirose. Phonetica, in press.

Auditory and Linguistic Processes in the Perception of Intonation Contours. M. Studdert-Kennedy and K. Hadding. Language and Speech, in press.

*The Specialization of the Language Hemisphere. A. M. Liberman. Invited paper to appear in the proceedings of the Intensive Study Program of the Neurosciences Research Program, Boulder, Colo., July 1972. (Cambridge: MIT Press, in press).

Word-Final Stops in Thai. In Thai Phonetics and Phonology, ed. by R. B. Noss and J. Harris. (Bankok, in press).

The following four papers will appear in Current Trends in Linguistics, Vol. XII, ed. by Thomas A. Sebeok. (The Hague: Mouton, in press):

Phonetics: An Overview.
Arthur S. Abramson.

The Perception of Speech.
Michael Studdert-Kennedy.

Speech Synthesis for Phonetic and Phonological Models.
Ignatius G. Mattingly.

On Timing in Speech.
Leigh Lisker.

Field Evaluation of an Automated Reading System for the Blind. P. W. Nye, J. D. Hankins, T. Rand, I. G. Mattingly, and F. S. Cooper. IEEE Transactions on Audio and Electroacoustics, in press.

*Reading and the Awareness of Linguistic Segments. Isabelle Y. Liberman, Donald Shankweiler, Bonnie Carter, and F. William Fischer

*A Preliminary Report on Six Fusions in Auditory Research. James E. Cutting.

*A Right-Ear Advantage in the Retention of Words Presented Monaurally. M. T. Turvey, David Pisoni, and Joanne F. Croog.

*Ear Advantage for Stops and Liquids in Initial and Final Position. James E. Cutting.

*A Right-Ear Advantage in Choice Reaction Time to Monaurally Presented Vowels: A Pilot Study. Michael Studdert-Kennedy.

*Visual Storage or Visual Masking?: An Analysis of the "Retroactive Contour Enhancement" Effect. M. T. Turvey, Claire Farley Michaels, and Diane Kewley Port.

*Hemiretinae and Nonmonotonic Masking Functions with Overlapping Stimuli. Claire Farley Michaels and M. T. Turvey.

Reports and Oral Presentations

Lectures on Constructivism. Ruth S. Day. State University of New York at Buffalo, 21-22 June 1972.

Temporal Order Judgment in Speech. Ruth S. Day. Presented to the Symposium on Serial Order in Behavior and Thought, sponsored by the Mathematical Social Science Board of the Center for Advanced Study in the Behavioral Sciences, Ann Arbor, Mich., 17-21 July 1972.

Implications of Cognitive Psychology for the Educational Process. Ruth S. Day. Invited address, North Central Association of Colleges, University of Denver, 9 August 1972.

*An Automated Reading Service for the Blind. J. Gaitenby, G. Sholes, T. Rand, G. Kuhn, P. Nye, and F. Cooper. Presented at the Carnahan Conference on Electronic Prosthetics, Lexington, Ky., 22 September 1972.

Phonetic Prerequisites for First-Language Acquisitions. I. G. Mattingly. Presented at the International Symposium on First-Language Acquisition, Florence, Italy, September 1972.

What Makes a Task "Verbal?" Ruth S. Day. Presented in symposium on Hemispheric Asymmetry: Stimulus Effect vs. Process Effect. American Psychological Association, Honolulu, Hawaii, September 1972.

The Stress Contrast Mechanism. K. S. Harris. A talk presented at the Linguistics Section meeting of the New York Academy of Sciences, 16 October 1972.

Colloquia. Ruth S. Day. Program in Psycholinguistics, University of Michigan, 13 July 1972; Department of Psychology, Princeton University, 4 October 1972; and Department of Psychology, University of Pennsylvania, 17 October 1972.

*A Continuum of Cerebral Dominance for Speech Perception. Michael Studdert-Kennedy and Donald Shankweiler. Read before the Academy of Aphasia, Rochester, N.Y., October 1972.

Laryngeal Control in Vocal Attack. H. Hirose and T. Gay. Presented at the Annual Meeting of the American Speech and Hearing Association, San Francisco, Cal., November, 1972.

*Velar Activity in Voicing Distinctions: A Simultaneous Fiberoptic and Electromyographic Study. Fredericka Bell-Berti and Hajime Hirose. Presented at the American Speech and Hearing Association Convention, San Francisco, Cal., November 1972.

*Electromyographic Study of Speech Musculature During Lingual Nerve Block. Gloria J. Borden, Katherine S. Harris, and Lorne Catena. Presented at the American Speech and Hearing Association Convention, San Francisco, Cal., November 1972.

*Stop Consonant Voicing and Pharyngeal Cavity Size. Fredericka Bell-Berti and Hajime Hirose. Presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., November 1972.

*Physiological Aspects of Certain Laryngeal Features in Stop Production. H. Hirose, L. Lisker, and A. S. Abramson. Presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., November 1972.

Techniques for Processing EMG Signals. D. K. Port. Presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., November 1972.

A Parallel Between Degree of Encodedness and the Ear Advantage: Evidence from an Ear-Monitoring Task. J. E. Cutting. Presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., November 1972. (Also in SR-29/30, 1972, as: A Parallel Between Encodedness and the Magnitude of the Right Ear Effect.)

*Electromyographic Study of the Tones of Thai. Donna Erickson and Arthur S. Abramson. Presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., November 1972.

*A Parallel Between Degree of Encodedness and the Ear Advantage: Evidence from a Temporal Order Judgment Task. Ruth S. Day and James M. Vigorito. Presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., 1 December 1972.

*Memory for Dichotic Pairs: Disruption of Ear Report Performance by the Speech-Nonspeech Distinction. Ruth S. Day, James C. Bartlett, and James E. Cutting. Presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., 1 December 1972.

*Perceptual Processing Time for Consonants and Vowels. David B. Pisoni. Presented at the 84th meeting of the Acoustical Society of America, Miami Beach, Fla., 1 December 1972.

Is Syntax Necessary? (In Phrasal Stress Assignment By Rule). J. H. Gaitenby. Presented at International Business Machine Corp., Thomas J. Watson Research Center, Yorktown Heights, N. Y., 19 December 1972.

APPENDICES

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers:

SR-21/22 to SR-29/30

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	
SR-28	October - December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	

Errata

SR-28 (October - December 1971)

The Activity of the Intrinsic Laryngeal Muscles in Voicing Control: An Electromyographic Study. Hajime Hirose and Thomas Gay (115-142).

Page 124, figure 5, lower bar should read: bʌzə
Page 126, table III, letters under Subject 2 should read: /p/ /t/ /k/

Page 138, paragraph 4, adductor should read: abductor
Page 141, line 2, statis should read: static

SR-29/30 (January - June 1972)

The Phi Coefficient as an Index of Ear Differences in Dichotic Listening. Gary M. Kuhn (75-82).

Page 76, lines 8 and 9 should read: L = the number of left-ear responses
R = the number of right-ear responses

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Haskins Laboratories, Inc. 270 Crown Street New Haven, Connecticut 06510	Unclassified 2b. GROUP N/A

3. REPORT TITLE

Status Report on Speech Research, no. 1/32, July-December 1972.

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Interim Scientific Report

5. AUTHOR(S) (First name, middle initial, last name)

Staff of Haskins Laboratories; Franklin S. Cooper, P.I.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
January 1973	254	337

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
ONR Contract N00014-67-A-0129-0001 NIDR: Grant DE-01774 NICHD: Grant HD-01994 NIH/DRFR: Grant RR-5596 NSF: Grant GS-28354 VA/PSAS Contract V101(134)P-71 NICHD Contract NIH-71-2420 The Seeing Eye, Inc.--Equipment Grant	SR-31/32 (1972) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.*

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
N/A	See No. 8

13. ABSTRACT This report (1 July-31 December) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Manuscripts cover the following topics:

-Specialization of the Language Hemisphere
-Cerebral Dominance for Speech Perception
-Degree of Encodedness and Ear Advantage
-Memory for Speech-Nonspeech Dichotic Pairs
-Ear Advantage for Stops and Liquids in Initial and Final Position
-Ear Advantage in Retention of Words Presented Monaurally
-Ear Advantage in Choice Reaction Time to Vowels Presented Monaurally
-Perceptual Processing Time for Consonants and Vowels
-Six Fusions in Auditory Research: A Preliminary Report
-Constructive Theory, Perceptual Systems, and Tacit Knowledge
-Hemiretinae and Nonmonotonic Masking Functions with Overlapping Stimuli
-Retroactive Contour Enhancement Effect: Visual Storage or Visual Masking?
-Reading and the Awareness of Linguistic Segments
-Machines and Speech
-Automated Reading Service for the Blind
-Physiological Aspects of Certain Laryngeal Features in Stop Production
-Effect of Speaking Rate on Labial Consonant Production: EMG and Cine Study
-Stop Consonant Voicing and Pharyngeal Cavity Size
-EMG Study of Speech Musculature During Lingual Nerve Block
-Velar Activity in Voicing Distinctions: EMG and Endoscopic Study
-Electromyographic Study of the Tones of Thai

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6811

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

UNCLASSIFIED
Security Classification

A-31408

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech coding						
Speech perception						
Ear advantage						
Dichotic listening						
Speech processing						
Auditory fusion						
Perceptual systems						
Visual masking						
Reading: phonetic awareness						
Speech input to computers						
Reading machines for the blind						
Physiology of speech production						
Electromyography of speech production						
Visual information processing						